Julius-Maximilians-Universität Würzburg

Professur für Inverse Probleme

Bayesian Inference for Composition of Hypotheses in Sequential Data

Datum:
28. September 2024

Masterarbeit von
Max Johann Levermann

Betreuer:
Prof. Dr. Frank Werner
Prof. Dr. Andreas Hotho

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Würzburg, 28. September, 2024

Max Johann Levermann

Contents

1

Introduction

Understanding human decision making is a research area that touches upon many disciplines even outside the scope of computer science and mathematics. However, having statistical tools and methods to precisely formulate problems and analyse behavioural properties, is common ground for all scenarios where decision making for humans is involved. In this work we will be particularly interested in the analysis of navigational behaviour and way finding of humans. These problems are often modelled through complex networks, which give an underlying structure for admissible trails that can be taken throughout the navigation process. A trail simply represents a sequence of states that have been traversed by a human through a given network. For our analysis, data will always be given in the form of trails and they are therefore an important concept for our purposes. Human trails appear in many different forms, even in our every day life, such as navigating through the web or finding the shortest way to the next supermarket. Even though these two scenarios do not have much in common, both can be modelled using a network which inherits the most integral structural properties. For instance, a web graph with hyper links and a city map with determined locations connected by streets would be adequate network representations for the two scenarios, respectively.

Analysing human navigation through such networks by formulating and comparing hypotheses that represent different beliefs about transitional behaviour is the central motivation of the HypTrails approach [1]. This approach utilizes a first order Markov chain to model transitions which are in turn used to calculate the likelihood of observed trails. The transition probabilities used for the Markov chain and the likelihood calculation are taken from different hypotheses, which itself yield different reasoning for navigational behaviour. Actually, the HypTrails approach leaves the frequentist realm of simple likelihood calculations by modelling different hypotheses as Dirichlet priors and interpreting the transition probabilities of the Markov chain as a categorical distribution, which, following the Bayesian paradigm, allows for the computation of evidence values (marginal likelihood). With these values, Bayes factors can be determined and a partial ordering of all hypotheses can be established.

A natural extension to this approach is given by MixedTrails [2]. The MixedTrails framework introduces group assignment probabilities, which allow observed transitions to be explained by different hypotheses. With this, not only single hypotheses can be compared, but also compositions of hypotheses. MixedTrails directly builds upon HypTrails as it utilizes all its theoretical elements and generalizes them to the formulation of mixed hypotheses. We will thoroughly present both approaches in a later chapter. An open question that remains to be answered and has been posed by the authors of the mentioned works [1], [2], aims at determining optimal compositions of hypotheses. That is, for a given data set and a given set of hypotheses, what combination of these hypotheses explains the data best. We are going to propose a solution to this question, that relies on performing Bayesian inference over the space of mixing ratios.

1.1 Bayesian Inference via Markov Chain Monte Carlo

Loosely speaking, Bayesianism refers to the interpretation of probability as a degree of belief with respect to given observations and prior knowledge. Under this paradigm, Bayesian inference describes the process of inferring parameters of a certain model by incorporating the two quantities - prior and data - into the process. That is, the prior belief about the occurrence of an event is updated according to data that has been observed. At the core of this updating process lies Bayes theorem, which connects the prior to the posterior. The posterior is the object of interest, as it yields the updated belief after data has been observed.

Bayesian inference is the first developed statistical inference method. It goes back to Thomas Bayes, who formulated a first version of the theorem, which got published three years after his death, in 1764. After some years of downswing, Bayesian inference gained new popularity when the connection to Markov Chain Monte Carlo methods was made in the 1990s. The fusion of these two theories was supplemented with the presence of a growing computing capacity of processors in that time, which could fulfil the demand of computationally expensive Monte Carlo simulations.

Monte Carlo approaches are randomized algorithms that simulate probability distributions by generating samples as approximations. A large class of such algorithms is covered under the collective term of Markov Chain Monte Carlo (MCMC) methods. As the name suggests, these methods utilize a Markov Chain for their generation of states and samples. Arguably, the best known of such MCMC methods, is the Metropolis Hastings algorithm. This algorithm is able to generate samples from distributions that are only known proportionally, which is the case for posterior distributions. Essentially, to sample from a distribution, the idea of MCMC is to combine the randomness aspect from Monte Carlo techniques with a (hopefully) informative proposal scheme that eventually becomes independent of the starting point of the Markov Chain that it is run upon.

1.1.1 Contributions

This thesis connects to the existing methodology about the comparison of hypotheses for sequential data and it utilizes the frameworks that were introduced in [1] and [2]. More particularly, the main interest of this work is to find optimal compositions of hypotheses by employing the Metropolis Hastings algorithm. One central aspect of this work is that the parameter space, which the Metropolis Hastings algorithm generates samples from, is directly linked to MixedTrails and their approach to the evidence calculation of mixtures of hypotheses. More precisely, the evidences are integrated into the sampling procedure as log likelihoods of the current parameter configuration, ranking the plausibility of the respective combination of hypotheses.

Our approach to find optimal compositions via the Metropolis Hastings algorithm is employed for different synthetic data sets demonstrating not only its theoretical functionality but also showcases its contributions to the posed open problem.

1.2 Structure of this Work

This thesis consists of three main chapters.

Chapter 2 covers all the background knowledge that this work builds upon and which is needed to formulate and understand the concept of composition of hypotheses. To this end, stochastic processes, specifically Markov chains on discrete state spaces are a fundamental concept. We further give a substantial summary of the HypTrails and the MixedTrails approach and the way that they utilize Bayes factors as a quantifier for the fit of different hypotheses.

Chapter 3 is devoted to a thorough introduction to our method and presents all the theory that is needed to prove fundamental convergence results for the Metropolis Hastings algorithm in an almost self-sufficient manner. We define Markov chains on general state spaces together with the desired property of stationarity. Then, we state the algorithm for our specific use cases and show that the transition kernel, which describes the Metropolis Hastings algorithm, indeed fulfils all requirements to converge to stationarity.

In chapter 4 experiments are carried out to show the applicability of our method. We start out by applying the algorithm to simple synthetic data sets and show that we are able to restore the generative settings of the data set. We then add noise to the data and analyse the behaviour of different hypotheses compositions leading to an approach were also the belief level of a hypothesis is integrated into the sampling procedure. Finally, we show how we can use the influence of the prior to be able to obtain and then solve a multi modal sampling problem with the Metropolis Hastings algorithm.

2

Background

This chapter is devoted to a thorough introduction to all concepts that are needed throughout this work. Section 2.1 starts with the most basic definitions concerning probability theory in general. After that, stochastic processes are defined, along with the two protagonists, random walks and Markov chains. Markov chains are a central tool that is used by HypTrails [1] and their Bayesian approach to the comparison of hypothesis for sequential data. In later chapters, we see how random walks are used to generate synthetic sequential data, that, in turn, is used by HypTrails. We close this chapter by summarizing the ideas to extend the approach taken by HypTrails to compare *compositions* of hypotheses, as established by [2] named MixedTrails.

2.1 Probability Theory

2.1.1 Probability Spaces

Probability theory at its core is used to mathematically describe the outcome of some experiment or observation where randomness is involved. The set of all possible outcomes is called the sample space and is denoted as Ω. For example, $\Omega = \{1, 2, 3, 4, 5, 6\}$ when throwing a dice. Secondly we want to assign probabilities to outcomes. For a fair dice the probability to roll any number is given by 1/6. The function responsible for this is called the probability measure $P : \mathcal{P}(\Omega) \to [0, 1]$. In general, the probability measure does not need to be defined on the whole power set of Ω but on a smaller subset, the so called σ-algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$. A σ-algebra needs to contain the whole space Ω, needs to be closed under complements as well as under countable unions. Elements of $\mathcal{A}$ are called events. We can now model the probability of the event "rolling an uneven number" as $P(\{1, 3, 5\}) = 1/2$. The calculation here is simple but we actually used the so called σ-additivity of the measure P. That is, for any countable union of pairwise disjoint sets the overall probability is given by the sum of the probabilities of the individual sets. In our example, this translates to $P(\{1, 3, 5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = 1/6 + 1/6 + 1/6$. Lastly, we require that $P(\Omega) = 1$ as well as $\Omega \neq \emptyset$. With all the discussed properties we call the tuple $(\Omega, \mathcal{A}, P)$ a probability space. Leaving out the measure, the tuple $(\Omega, \mathcal{A})$ is simply called a measurable space.

2.1.2 Random Variables and Probability Distributions

For a given probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and a measurable space $(E, \mathcal{B})$ we call a function $X : \Omega \to E$ a random variable, if it is $(\mathcal{A} - \mathcal{B})$-measurable, i.e. $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$. This property allows us to naturally define a probability measure $\mathcal{P}_X$ on E as $\mathcal{P}_X(B) = \mathcal{P}(X^{-1}(B)) = \mathcal{P}(\{w \in \Omega : X(w) \in B\}) =: \mathcal{P}(X \in B)$ for any $B \in \mathcal{B}$ which is called the (probability) distribution of X. In general, it is not necessary to have a random variable to define a probability distribution. Any probability measure P within a probability space $(\Omega, \mathcal{A}, P)$ already defines a probability distribution, but mostly, they come together with a random variable. For a discrete space Ω, P is referred to as probability mass and for a continuous state space as probability density. Therefore, a probability distribution is fully described by P. We are going to introduce two examples of probability distributions, that will be used in later chapters.

Exponential Distribution The probability density function of the exponential distribution $Exp(\lambda)$ is defined for $x \geq 0$ and is given by

$$f_\lambda(x) = \lambda e^{-\lambda x}.$$

This distribution has the interval $[0, \infty)$ as its support. The parameter $\lambda \in \mathbb{R}_{>0}$ is referred to as the *rate* of the distribution. The expected value is given by $1/\lambda$.

Dirichlet Distribution The Dirichlet distribution, denoted as $Dir(\alpha)$ is a continuous probability distribution with parameter vector $\alpha \in \mathbb{R}^k_{>0}$. It is a multivariate generalisation of the Beta distribution with support given by the $(k-1)$ dimensional standard simplex $\Delta \subset [0,1]^k$, i.e. $\sum_{i=1,\ldots,k} x_i = 1$ for any $x = (x_1, \ldots, x_k)$ with positive probability. The parameter vector α influences the probability density, by assigning more probability mass to regions that correspond to large entries of the vector α. That is, each entry of α corresponds to one corner of the underlying simplex and more probability mass is shifted towards those corners that have large values. If all entries of the parameter vector are greater or equal than 1, then there is a single mode in the probability density, whose location is a compromise between the magnitude of the entries, but ultimately is closest to the largest entry. A special case is given by $\alpha = (1, \ldots, 1)$, which yields a uniform distribution over the simplex. More generally, if all entries in the parameter vector are the same, we obtain a symmetric Dirichlet distribution, for which the probability mass concentrates in the center of the simplex (entries larger than 1) and a higher concentration is given for larger values. If all entries of α lie in the open interval $(0,1)$, the density gets inverted, in the sense that now there are modes in every corner of the simplex. The modes becomes larger the closer an entry lies to 0. Overall, in this scenario, points on the simplex that are sparse, i.e. that have mainly zero entries, have the most assigned probability mass. Finally, the probability density function is given by

$$f_\alpha(x_1, \ldots, x_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1},$$

where $B(\alpha)$ denotes the multivariate Beta function, which itself is defined via the Gamma function Γ, an extension of the factorial function to real numbers, as

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}.$$

2.2 Stochastic Processes

So far we have considered single random variables on a probability space $(\Omega, \mathcal{A}, P)$ to model observations or outcomes from a random experiment. Stochastic processes are defined as a collection of random variables that are all defined on the same probability space. The random variables are indexed by some given set I. For our purposes we will restrict I to be subset of $\mathbb{R}_+$ which yields a time interpretation of the process. In general, arbitrary sets are admissible.

Definition 2.1 (stochastic process).
Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(S, \mathcal{B})$ a measurable space and $I \subseteq \mathbb{R}_+$ an index set. A family $X = (X_t)_{t \in I}$ of random variables $X_t : \Omega \to S$ is called a stochastic process with state space S.

If not stated otherwise $\mathcal{B} = \mathcal{B}(S)$, denoting the Borel sigma algebra.

Being able to uniquely address each random variable using their index, enables us, to compare any selection of different states associated with their respective random variable. For example, an *increment* computes the difference of two instantiations of the process.

Characterization Dependent on properties of the index set I and the state space S we can characterize different types of stochastic processes.
For a discrete index set I we refer to the stochastic process as time discrete. The most important case is $I = \mathbb{N}_0$ which simply is a sequence of random variables $(X_n)_{n \in \mathbb{N}_0}$. We can further split time discrete processes into those with a discrete or continuous state space, i.e. the random variables of the process having a discrete or continuous distribution. For our cases the most important example of time discrete stochastic processes with a discrete state space are Markov chains, which will be discussed in detail in the following section.
For a continuous index set I the process is called time continuous. A prominent example for a time continuous process with a continuous state space is the Wiener process which mathematically models the Brownian motion. Poison processes with $S = \mathbb{N}_0$ are a typical example for time continuous processes with a discrete state space.

2.2.1 Random Walks

Random walks are arguable among the most prominent examples of time discrete processes and they will also be of particular interest for our purposes. We will mainly use them to generate transitional or sequential data. As this type of data sits at the core of all the works that this thesis builds upon, it also plays an important role in the upcoming chapters.

Random Walk on $\mathbb{R}^d$ Formally, for a sequence of i.i.d. random variables $X_1, X_2, \ldots$ that take values in $\mathbb{R}^d$, and

$$Z_n = \sum_{i=1}^{n} X_i,$$

we call the stochastic process $(Z_n)_{n\in\mathbb{N}}$ a random walk on $\mathbb{R}^d$. In each time step an increment is determined by the underlying probability distribution on $\mathbb{R}^d$. If $P(X_i = 1) = 1/2$ and $P(X_i = -1) = 1/2$ and the process is defined on $\mathbb{Z}$ instead of $\mathbb{R}^d$, we call (Z_n) a one dimensional symmetric random walk.

Random Walk on a Graph A random walk can also be defined on a graph. The involved random variables are then no longer i.i.d., because each step is taken according to the current node and the existing graph structure. To obtain a random walk on a graph, the process is initialized at any node. Then, in each step, the next state is chosen uniformly to be one of neighbouring nodes. If the edges of the graph are weighted, then the next state is chosen according to the induced probability distribution from the weights that connect the current node to its neighbours. If we further assume that nodes can inherit different values or features, than the next state can also be chosen under consideration of these features. A random walk on a graph outputs a sequence of nodes, which we sometimes refer to as a trail.

2.2.2 Markov Chains on a Discrete State Space

As we have already mentioned, Markov chains are a special case of stochastic processes. They can be defined on both, a continuous or a discrete state space S. For the purposes of this chapter and the theory that is derived in sections 2.3.2 and 2.3.3, we shall be concerned with the latter. A matrix $\mathbf{p} \in \mathbb{R}^{S\times S}$ is *stochastic*, if all its entries are positive and the rows sum up to 1:

(i) $p_{ij} \geq 0$ for all $i, j \in S$

(ii) $\sum_j p_{ij} = 1$.

With these two properties each row of the matrix can be interpreted as a discrete probability distribution over S.

Definition 2.2 (Markov chain).
Let S be a countable set, $\mathbf{p} = p_{ij} \in S \times S$ a stochastic matrix and π_0 some distribution on S. A stochastic process $X = (X_t)_{t\in\mathbb{N}_0}$ is called a $(\pi_0, \mathbf{p})$-Markov chain, if

(i) $P(X_0 = i) = \pi_0(i)$, $i \in S$,

(ii) for all $t \in \mathbb{N}_0$ and $i_0, \dots, i_{t+1} \in S$ with $P(X_0 = i_0, \dots, X_t = i_t) > 0$ it holds that

$$P(X_{t+1} = i_{t+1}|X_0 = i_0, \dots, X_t = i_t) = P(X_{t+1} = i_{t+1}|X_t = i_t),$$

(iii) for all $t \in \mathbb{N}_0$ *and* $i_t, i_{t+1} \in S$ *with* $P(X_t = i_t) > 0$ *it holds that*

$$P(X_{t+1} = i_{t+1} | X_t = i_t) = p_{i_t i_{t+1}}.$$

The last property connects the stochastic process to the stochastic matrix. The probability of transitioning from any state $i \in S$ to any state $j \in S$ is precisely given by the the entry p_{ij}. Note that this transition probability is independent of the time t and we could reformulate the condition as $P(X_1 = j | X_0 = i) = p_{ij}$. A Markov chain that doesn't satisfy this property, meaning that the transition probabilities between states are changing over time, is called *inhomogen*. The second property is the central characteristic of a Markov chain and is called the memoryless property. This notion is due to the fact, that the probability of transitioning to the next state is independent of the past states as they are "forgotten" by the process.

2.3 Hypothesis Driven Analysis of Sequential Data

In this section, we are going to give a detailed introduction to some fundamental concepts that are needed to analyse and compare hypotheses for sequential data. First, we are going to elaborate on properties and characteristics of sequential data, as well as looking at some explicit data sets that will be used throughout this work. Next, we are going to introduce the HypTrails framework [1] which allows us to compare hypotheses utilizing the Bayesian paradigm. This is followed by the MixedTrails approach [2] which is a natural extension of HypTrails, that allows to combine hypotheses and compare these different compositions. Compositions of hypotheses, specifically finding an optimal composition, will play a central role in this work.

2.3.1 Sequential Data and Human Navigation

Sequential data doesn't really have a common definition, the term rather expresses the existence of some kind of dependency or order within the data. This dependency of data points can exist for different reasons. In the field of natural language processing for example, where data mainly consists of sentences, the dependency of words is given by the structure of a sentence, which itself follows grammar rules of the language. A different example is the analysis of time series data, where data points are samples from some system at different points in time. To make future predictions or to find anomalies in the series, one tries to exploit the time dependency between data points. As a more general approach to sequential data (over a discrete state space) we can think of data points as nodes in a graph and the dependencies of different points are given by the edges in the graph. Of course, edges can be directional or weighted as well. Examples of graph structured data are citation networks or web graphs where nodes represent articles or web pages and edges indicate citations or hyperlinks, respectively. Nodes can also inhibit different values for categories of interest, for example author or publication year in the case of the citation network. Data that has an underlying non trivial graph structure is referred to as a complex network.

Analysing human navigation through such complex networks is one of the main motivations of the HypTrails approach, which we will discuss in the upcoming section 2.3.2. Human navigational data, or human trails, exist everywhere, where human decision making is involved. Examples are real world travel sequences over locations, such as way finding in a city, where states are given by certain locations. Another example is web navigation, where users click through websites, where hyperlinks follow an underlying web graph and a states are given by the nodes of the web graph.

2.3.2 HypTrails

HypTrails is a tool, that allows to compare different hypotheses, which aim to explain navigational or, more generally, sequential data. A hypothesis expresses a belief about how an observed transition between any two states can be explained. To obtain hypotheses, we consider different driving factors, that could generate the observed transitions. For example, when navigating through a city with a certain destination, driving factors could be coupled to the used transportation type, as driving by car will yield different navigational behaviour as riding by bike or as walking. Generally, to be able to determine hypotheses in arbitrary domains might require some prior knowledge within that domain.

Representing Beliefs For a discrete and finite state space $S = \{s_1, \dots s_n\}$, we model certain beliefs about transitional behaviour via a stochastic matrix $\phi \in \mathbb{R}^{n \times n}$. An entry ϕ_{ij} represents the probability to observe a transition from state s_i to s_j. Each row of this matrix sums to 1, so we can think of a row as a (discrete) probability distribution over the state space S. To incorporate the belief into the matrix, higher probabilities are assigned to entries that match well with the belief and low probabilities are assigned to entries that are contrary to the belief. Therefore, different beliefs correspond to different stochastic matrices.

Markov Chain Modelling The HypTrails approach utilizes a first order Markov chain model to describe the transitional behaviour of different observed trails, given by $D = \{t_1, \ldots, t_m\}$. That is, the probability for any transition t_k between states s_i and s_j, is given by $P(X_{n+1} = s_j \mid X_n = s_i, \ldots, X_0 = s_0) = P(X_{n+1} = s_j \mid X_n = s_i)$, for a Markov chain (X_n) defined on the state space S. Due to the Markov property, we can summarize these probabilities in a transition matrix: $P(X_{n+1} = s_j \mid X_n = s_i) = \phi_{ij}$. This is precisely the point, where our belief, represented by a stochastic matrix, can be connected to the Markov chain.

For us, a data set consists of observed transitions. For a Markov Model the likelihood for a given data set D is calculated as

$$P(D \mid \phi) = \prod_{t_k \in D} \phi_{i_k j_k} = \prod_{s_i, s_j \in S} \phi_{ij}^{n_{ij}},$$

where n_{ij} denotes the number of observed transitions between states s_i and s_j.

Bayesesian Inference Bayesian inference is used to infer the unknown parameter θ of some model from observed data D. In this process, the model and its parameters, as well as the observed data are understood to be random variables and therefore, follow some probability distribution. Furthermore, a prior distribution is imposed over the parameter space. For a given likelihood function, it is always desirable to choose a *conjugate* prior, because this implies, that the posterior belongs to the same probability distribution family as the prior. Bayes theorem, stated in the following, illustrates the mentioned dependencies. By H, we denote some hypothesis or model representing a belief. A hypothesis is going to be expressed via the prior probability distribution $P(\theta \mid H)$, yielding a posterior in the form of

$$P(\theta \mid D, H) = \frac{P(D \mid \theta, H) P(\theta \mid H)}{P(D \mid H)}.$$

In this formula, $P(D \mid \theta, H)$ is the likelihood of observed data given parameters θ and hypothesis H. We can think of the observed data to be arranged in a transition count matrix of size $n \times n$ and the entry (i, j) corresponds to the number of counts from state s_i to s_j, denoted as n_{ij}. We denote the i'th row of this transition count matrix as n_i (we will be able to distinguish this variable from the number of states n). Every observed transition from this row can be interpreted to follow a categorical distribution $Cat(n, \theta_i)$ with unknown parameter vector θ_i. This parameter vector, that shall model the observed data, is obtained from the prior distribution, according to Bayes formula. HypTrails uses a Dirichlet distribution as a prior, which is a conjugate prior of the Categorical distribution.

In the following, we are going to discuss how to incorporate hypotheses, corresponding to some belief, into the prior, i.e. we discuss how to obtain $P(\theta \mid H)$. This process is called elicitation, and we give a reduced version of the trial roulette method used in [1]. A belief can be expressed as a stochastic matrix ϕ, which itself is coupled to the Markov chain model describing the transitional behaviour between states. To obtain a Dirichlet prior, we again interpret the matrix ϕ row wise, that is, we define a Dirichlet prior $Dir(\alpha_i)$ for each state s_i. The entries of α_i need to be strictly positive. They are obtained as follows. First, add a 1 to all entries to avoid zero parameters. Then, a concentration parameter $\kappa \in \mathbb{R}_{\geq 0}$ is introduced to up scale all entries. The calculation for the parameter vector of the Dirichlet distribution associated with state s_i now reads as

$$\alpha_i = \kappa \cdot \phi_{s_i} + 1, \tag{2.1}$$

where ϕ_{s_i} is the i'th row of ϕ.

The concentration parameter κ plays a crucial role in the analysis and comparison of hypotheses. The larger we choose κ the more belief is put into the hypothesis. This is simply due to the way the Dirichlet distribution behaves. The larger we choose the entries, the more probability mass is concentrated around the initially incited probabilities of the belief matrix ϕ. For a thorough analysis, we will not restrict to a single value κ but use a range of different values.

Since the Dirichlet distribution is a conjugate prior with respect to the Categorical distribution, the posterior follows a Dirichlet distribution as well. More precisely the following holds. Let α_i be the parameter vector for the Dirichlet distribution with respect to some fixed concentration parameter κ and belief matrix ϕ. Then

$$\begin{aligned} \theta_i \mid \alpha_i = (\theta_{i1}, \dots, \theta_{in}) &\sim Dir(\alpha_i) \\ D_i \mid \theta_i = (dst_k)_{k \in |D_i|} &\sim Cat(\theta_i) \end{aligned}$$

and

$$\theta_i \mid D_i, \alpha_i \sim Dir(\alpha_{i1} + n_{i1}, \dots, \alpha_{in} + n_{in}),$$

where D_i denotes the i'th row of the transition count matrix, so it contains all transitions that belong to state s_i. More precisely, the j'th entry of

D_i contains the number of counts of an observed transitions from state s_i to state s_j. Since the Dirichlet distribution is a conjugate prior of the categorical distribution, we obtain the parameter vector of the Dirichlet posterior - row wise - as the sum of the actual observed transitions n_i and α_i. Due to this relation, we will refer to the entries of α_i as pseudo counts.

Bayesesian Model Comparison So far, we have seen how we can express a hypothesis as a prior and have further established the dependencies between prior, likelihood and posterior. Ultimately, the goal is to compare different hypothesis to be able to determine which hypothesis is the best fit for the given data set. To do this, we use the marginal likelihood $P(D \mid H)$, which expresses the probability of the data given the hypothesis H. We can not simply use the marginal likelihood as an absolute measure for the plausibility of a hypothesis, because the value is dependent on the size of the data set as well as the pseudo counts. Instead, we use Bayes factors, which measure the plausibility of a hypotheses relative to another. That is, for two hypothesis H_1 and H_2, the Bayes factor is given by

$$B_{1,2} = \frac{P(D \mid H_1)}{P(D \mid H_2)}.$$

The core aspect is, that we can only judge if a hypothesis explains the data better than another hypothesis, but can not judge if a hypothesis is generally a good fit for the data set. For a Bayes factor that is close to 1, we can consider two hypotheses to be equal. How close this needs to be, has to be decided by determining the significance level. For this, we, as well, refer to Kass and Rafterys interpretation table [3].

Calculating the Marginal Likelihood To close this section, we show how the marginal likelihood can be calculated. Since the probability density of the Dirichlet distribution integrates to 1, we get

$$\begin{aligned} 1 &= \int P(\theta_i \mid H) d\theta_i = \int \frac{1}{B(\alpha_i)} \prod_j \theta_{ij}^{\alpha_{ij}-1} d\theta_i \\ \Leftrightarrow B(\alpha_i) &= \int \prod_j \theta_{ij}^{\alpha_{ij}-1} d\theta_i. \end{aligned}$$

Now the marginal likelihood for a single row is given by

$$\begin{aligned}
P(D_i \mid H) &= \int P(D_i \mid \theta_i, H) P(\theta_i \mid H) d\theta_i \\
&= \int \prod_j \theta_{ij}^{n_{ij}} \frac{1}{B(\alpha_i)} \prod_j \theta_{ij}^{\alpha_{ij}-1} d\theta_i \\
&= \frac{1}{B(\alpha_i)} \int \prod_j \theta_{ij}^{n_{ij}+\alpha_{ij}-1} d\theta_i \\
&= \frac{B(\alpha_i + n_i)}{B(\alpha_i)}.
\end{aligned}$$

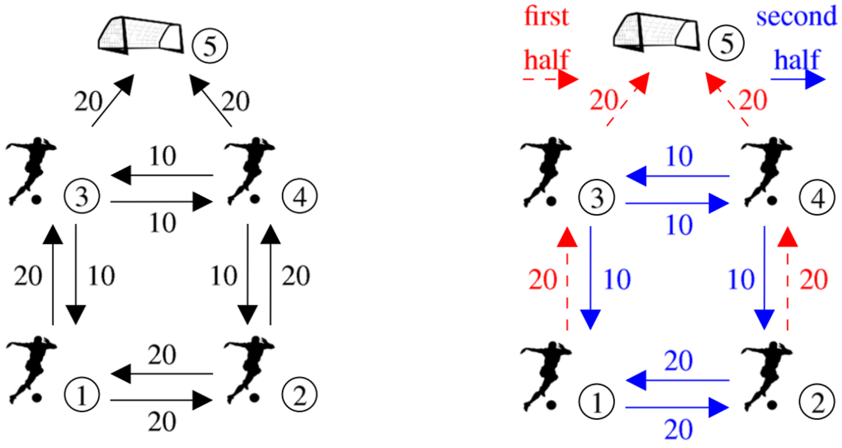
Since all rows are independent, the marginal likelihood for the whole data set is given by the product probability

$$P(D \mid H) = \prod_i \frac{B(\alpha_i + n_i)}{B(\alpha_i)}.$$

2.3.3 MixedTrails

The approach taken by MixedTrails [2] extends the HypTrails framework presented in the last section. The simple Markov chain model that is utilized by HypTrails is not able to explain heterogeneity in sequential data. The goal for MixedTrails is to tackle this problem by allowing for the formulation of more complex hypotheses. More precisely, mixed hypotheses are introduced. One such mixed hypothesis consists of different groups, where each group represents a different belief. An observed transition can be assigned to any of those groups. Therefore, different parts of the dataset can have different explanations. Furthermore, the process of assigning transitions to groups does not need to be deterministic but can also follow some (discrete) probability distribution. Finally, the approach taken by MixedTrails allows us to compare and rank different (mixed) hypotheses on sequential data, in the same manner as HypTrails. The marginal likelihood computation follows the *Mixed Transition Markov Chain* (MTMC) model, which generalizes the simple first order Markov chain utilized by HypTrails and allows for the evidence calculation of mixed hypothesis.

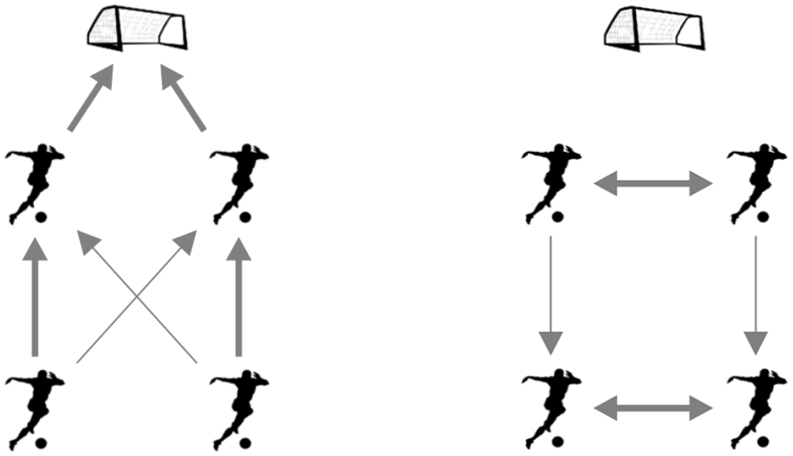
Heterogeneous Data



(a) Transitions (b) Heterogeneity within the Transition

Figure 2.1: Soccer Example taken from MixedTrails [2]

Heterogeneity in data can exist in many different scenarios. An easy way to recognize heterogeneity in data, is, if it can be split into disjoint parts, where each part is best explained by a different hypothesis. If we can identify those different parts, then we can deterministically assign each part to the corresponding hypothesis. An example for this situation is given by figure 2.1, where transitions are given as passes between soccer players or shots on a goal. States in this soccer example are given by the different positions and by the goal with given numbers 1 to 5. The numbers next to the arrows represent the number of observed transitions between the indicated states. In this example, a clear distinction between the first and the second half of the match can be made with regards to the strategy of the team. Looking at figure 2.1b we find, that in the first half, the strategy seems to be more offensive, as the observed transitions exclusively correspond to passes to a position that is closer to the goal. While in the second half, a more defensive strategy would explain the observed transitions better, as there are no passes played to an offensive position. The idea is now to formulate an offensive hypothesis for the first and a defensive hypothesis for the second half, as in figure 2.2 and assign all transitions from the first half deterministically to the first hypothesis and do the same for the second half and the defensive hypothesis.



(a) Offensive Hypothesis (b) Defensive Hypothesis

Figure 2.2: Different Hypotheses

Mixed Hypotheses

Let the state space again be given by $S = \{s_1, \dots, s_n\}$ and the observed transitions by $D = \{t_1, \dots, t_m\}$. A mixed hypothesis consists of a collection of belief matrices $\phi = \{\phi_1, \dots, \phi_o\}$ called the group transition probabilities, together with m - possibly different - stochastic vectors, collected in $\gamma = \{\gamma_t \mid t \in D\}$, called the group assignment probabilities. For any transition $t \in D$, the vector γ_t corresponds to a probability distribution over a set of groups $G = \{g_1, \dots, g_o\}$, where the number of groups is given by the number of belief matrices in ϕ, i.e. $\gamma_t = (\gamma_{g_1|t}, \dots, \gamma_{g_0|t})$. An assignment of transition t to group g means, that t is explained by the belief expressed in ϕ_g. This becomes relevant in the calculation process of the marginal likelihood, which we get from the MTMC model. For $\phi_g \in \phi$, we denote by $\phi_{ij|g}$ the probability to transition from state s_i to state s_j within the group g. We can see that a simple hypothesis, as used in HypTrails, can be obtained by assigning all transitions deterministically to one single group.

For the soccer example, we would split the data set into the first and second half, i.e. $D = D_1 \cup D_2$. Then two different group transition probability matrices ϕ_1 and ϕ_2 are constructed to express offensive and defensive transition behaviour, respectively. Now, any transition that has been observed in the first half $t \in D_1$ will have the group assignment probability $\gamma_t = [1, 0]$, so that it is assigned to the offensive strategy expressed in ϕ_1. Any transition that is observed in the second half $t \in D_2$ gets $\gamma_t = [0, 1]$ as a group assignment probability, so that these transitions are explained by the defensive strategy expressed in ϕ_2. Note that this example represents the special case of deterministic group assignments.

Mixed Transition Markov Chain (MTMC)

So far, we have seen how a mixed hypothesis can be formulated. Essentially, group assignment probabilities are added, that allow for a more nuanced explanation of transitions. We also saw, how HypTrails uses Bayesian inference with an hypothesis-integrated prior to calculate evidences and compare different - single - hypotheses. We are now going to introduce the Mixed Transition Markov Chain model, which will allow us to calculate the evidence of mixed hypotheses as well.

The group assignment probabilities γ from the mixed hypothesis are taken as the parameters of a categorical distribution. So, for each observed transition $t \in D$, we get a group assignment according to $Cat(\gamma_t)$. Then, within that group g, the transition t is explained by the belief expressed in ϕ_g in a HypTrails manner, following a simple first order Markov chain model that we saw in the last section. That is, the parameters $\theta_{i|g}$ are drawn from a Dirichlet distribution with parameters $\alpha_{i|g} = (\alpha_{i1|g}, \dots, \alpha_{in|g})$.

The following holds for every group $g \in G$ and every state $s_i \in S$

$$\theta_{i|g} \mid \alpha_{i|g} = (\theta_{i1|g}, \dots, \theta_{in|g}) \sim Dir(\alpha_{i|g}) \tag{2.2}$$

$$g \mid \gamma = (g_t)_{t \in |D_i|} \sim Cat(\gamma_t) \tag{2.3}$$

$$D_{i|g} \mid \theta_{i|g} = (dst_k)_{k \in |D_{i|g}|} \sim Cat(\theta_{i|g_k}) \tag{2.4}$$

and

$$\theta_{i|g} \mid D_{i|g}, \alpha_{i|g} \sim Dir(\alpha_{i1|g} + n_{i1|g}, \dots, \alpha_{in|g} + n_{in|g}),$$

where $D_{i|g}$ stands for all individual observed transitions from state s_i to any other destination state, given that this observed transition is assigned to group g. These transitions are then summed up as the transition counts within group g, denoted by $n_{i|g} = (n_{i1|g}, \dots, n_{in|g})$. Before we derive how to calculate the marginal likelihood for this model, we pick up on the topic of prior elicitation, as we shall see, that we require a different formula to determine the pseudo counts when dealing with probabilistic group assignments.

Prior Elicitation

To express a mixed hypothesis and make it fit into the Bayesian set up, that we introduced in 2.3.2, we will again use a Dirichlet distribution as a prior with parameter vector α. For a mixed hypothesis, we also have group assignment probabilities, so we need to take these into consideration, as well. For the special case of deterministic group assignments, that is, each transition in the data set is precisely assigned to one group, the elicitation process simply transfers over from HypTrails. Given some concentration parameter κ, for each state s_i and each group $g \in G$, the parameter vector is given by

$$\alpha_{i|g} = \kappa \cdot \phi_{s_i|g} + 1.$$

Again, 1 is added to each entry to avoid zeros and to obtain a well defined Dirichlet distribution.

If a mixed hypothesis uses probabilistic group assignments, it is necessary to account for all possible outcomes following the respective probability distribution for each transition. That is, we obtain $\alpha_{i|g}$ by not only looking at all observed transitions $t \in D$ and their probability to be assigned to group g, denoted by $\gamma_{g|t}$ but also at all other groups $g' \in G$ they can be assigned to with respective probability $\gamma_{g'|t}$. Overall, the following holds true:

$$\alpha_{i|g} = \kappa \cdot \left(\frac{1}{Z_i} \cdot \sum_{t \in D} \gamma_{g|t} \cdot \sum_{g' \in G} \gamma_{g'|t} \cdot \phi_{s_i|g'} \right) + 1.$$

We can think about $\gamma_{g|t}$ as the probability that the transition t is assigned to the group that it belongs to (as indicated by $\alpha_{i|g}$). Then, we assign for all actually possible group assignments g', by weighing the utilized belief $\phi_{s_i|g'}$ for that case with the probability that this assignment occurs $\gamma_{g'|t}$. For each state s_i a normalizing factor Z_i is needed, so that the underlying transition probabilities sum up to 1.

Marginal Likelihood for Mixed Transition Markov Chain Models

We are now going to show how the marginal likelihood for the Mixed Transition Markov Chain model can be calculated. In 2.3.3 we derived the MTMC model and saw how we can integrate groups and group assignment probabilities into the process. In particular, equation 2.4 is the only extension compared to the simple first order Markov chain model utilized by HypTrails. This is why, for a fixed group, we essentially get the same formula as for HypTrails:

$$\begin{aligned} P(D_{i|g} \mid H) &= \int P(D_{i|g} \mid H, \theta_{i|g}) P(\theta_{i|g} \mid H) d\theta_{i|g} \\ &= \int \left(\prod_j \theta_{ij|g}^{n_{ij|g}} \right) \left(\frac{1}{B(\alpha_{i|g})} \prod_j \theta_{ij|g}^{\alpha_{ij|g}-1} \right) d\theta_{i|g}. \end{aligned}$$

We now also need to account for all possible group assignment, that are drawn according to their respective probability given by γ following a Categorical distribution

$$g \mid \gamma = (g_t)_{t \in D_i} \sim Cat(\gamma_t).$$

We denote the set of all possible group assignments for all transitions by Ω and for those transitions belonging to D_i by Ω_i. Formally, $\Omega_i = \{\{(t, g_t)_{t \in D_i}\}\}$, where $g_t \sim Cat(\gamma_t)$. Let $\omega \in \Omega_i$ be any fixed group assignment, then the probability for this particular instantiation is given by $p_\omega = \prod_{t \in D_i} \gamma_{g_t|t}$, where $\gamma_{g_t|t}$ denotes the probability of transition t to belong to group g_t.

When accounting for all possible group assignments, to calculate the likelihood, we weigh each possible instantiation with its probability and take the product over all existing groups, according to the Categorical distribution they are sampled from:

$$P(D_i \mid H, \theta_i) = \sum_{\omega \in \Omega_i} p_\omega \prod_{g \in G} \prod_j \theta_{ij|g}^{n_{ij|g,\omega}} \tag{2.5}$$

Here, $n_{ij|g,\omega}$ denotes the observed counts, that now not only differ for every group but also for every instantiation ω.

Since the prior is not dependent on the transition counts, we only need to account for all possible groups, which again are drawn categorically, yielding

$$P(\theta_i \mid H) = \prod_{g \in G} \frac{1}{B(\alpha_{i|g})} \prod_j \theta_{ij|g}^{\alpha_{ij|g}-1}. \tag{2.6}$$

Overall, for one fixed state s_i, we get

$$\begin{aligned} P(D_i \mid H) &= \int P(D_i \mid H, \theta_i) P(\theta_i \mid H) d\theta_i \\ &= \int \sum_{\omega \in \Omega_i} p_\omega \prod_{g \in G} \prod_j \theta_{ij|g}^{n_{ij|g,\omega}} \prod_{g \in G} \frac{1}{B(\alpha_{i|g})} \prod_j \theta_{ij|g}^{\alpha_{ij|g}-1} d\theta_i \\ &= \sum_{\omega \in \Omega_i} p_\omega \prod_{g \in G} \frac{1}{B(\alpha_{i|g})} \int \prod_j \theta_{ij|g}^{n_{ij|g,\omega}+\alpha_{ij|g}-1} d\theta_i \\ &= \sum_{\omega \in \Omega_i} p_\omega \prod_{g \in G} \frac{B(n_{i|g,\omega} + \alpha_{i|g})}{B(\alpha_{i|g})}. \end{aligned}$$

Again, since all the states s_i and their corresponding group assignments, transition and pseudo counts are all independent from another, we obtain the marginal likelihood for the whole data set as the product over all states, with respect the a given instantiation ω and group g:

$$P(D \mid H) = \sum_{\omega \in \Omega} p_\omega \prod_{g \in G} \prod_{i \in S} \frac{B(n_{i|g,\omega} + \alpha_{i|g})}{B(\alpha_{i|g})}. \tag{2.7}$$

3

Method

In Chapter 2 we have seen how Bayesian inference can be used to compute the likelihood of different hypotheses that are incorporated into priors allowing for a comparison of such with regard to the explainability of the data. The theory and results that will be derived in this chapter also aim at doing Bayesian inference, but in a more classical sense. Ultimately, the goal is to find optimal mixing ratios of hypotheses by imposing a distribution over the respective parameter space and using a Markov Chain Monte Carlo approach (MCMC) to generate posterior samples, which in turn can be used to estimate expected or maximum likelihood values.

Classical estimators such as the maximum likelihood estimator or the method of moments require independent and identically distributed samples which are not at hand by default. There is a number of well known Monte Carlo approaches to generate samples, such as direct simulation or rejection sampling. In our case direct simulation is not feasible because the target distribution to be sampled from is not known. Classical rejection sampling suffers from the same problem, but it can also be implemented when the target distribution is proportional to some other density, which itself is known up to constant, i.e. $\pi(dx) \propto l(x)f(dx)$, and $l(x) \leq C$. Here, f, l and $C \in [1, \infty)$ need to be explicitly known. When all these values are known, rejection sampling can work really well, especially when the bound C is tight. This is, however, not guaranteed and we need other methods for scenarios, where this is not the case. For the following collection of methods it is enough to have $\pi = C \cdot f$, i.e. to know the target distribution up to a constant. The constant itself does not need to be known.

Markov Chain Monte Carlo Approaches Markov Chain Monte Carlo methods follow a slightly different approach to generate samples. Let again π be the target distribution. The general goal of MCMC algorithms is to estimate integrals over some measurable function h in the form of

$$\pi(h) = \int_{\mathcal{X}} h d\pi(x) = E_\pi[h(X)].$$

Here, the second equality only holds true if π is a *probability* measure and X an associated random variable. The approach taken by simple Monte Carlo methods is to approximate this integral by

$$\hat{\pi}(h) = \frac{1}{n} \sum_{i=1}^{n} h(X_i),$$

where $X_1, \ldots, X_n \sim \mu$ are i.i.d. random variables. As already mentioned, the generation process of these variables can be difficult or not even not feasible. The idea of MCMC algorithms is to instead construct a Markov chain, that has π as a stationary distribution and to let the chain run for some time. Eventually, the chain will reach its stationary distribution and samples generated by the chain follow the invariant target distribution. We call the stage from start to where we consider the chain to have reached stationarity burn-in phase. Obviously, samples that are generated in this phase can be disregarded since they are not assumed to follow the target distribution. Let b be the number of burned samples, then an integral approximation is given by

$$\hat{\pi}(h) = \frac{1}{n} \sum_{i=b+1}^{b+n} h(X_i).$$

The burn-in time b should be chosen, such that the Markov chain is as close as possible to the stationary distribution.

Before we look at the Metropolis Hastings algorithm, which will be of particular interest for us, we are going to motivate how MCMC approaches and Bayesian inference go together.

Bayes Motivation In the Bayesian scenario we assume that all involved distributions are parametrized, so instead of a distribution we can simply think of some parameter $\theta \in \mathbb{R}^d$ of a statistical model that we want to estimate. According to Bayes formula, given some observed data D the posterior distribution of the parameter θ is given by

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)},$$

where $P(\theta)$ is the prior distribution of θ, $P(D|\theta)$ is the likelihood of the observed data D under the parameter θ and $P(D)$ is the marginal likelihood of the data. While the prior distribution can be chosen according to prior beliefs, the marginal likelihood, given by $\int_{\mathcal{X}} P(D|\theta)P(\theta)d\theta(x)$, is often not known, since it is impossible to compute. As mentioned before, this is not a problem for MCMC methods, such as the Metropolis Hastings algorithm. They only require the target distribution to be known up to constant, which is the case here: $P(\theta|D) \propto P(D|\theta)P(\theta)$.

3.1 General State Space Markov Chains

In this section we are going to establish convergence results for Markov chains on a general state space. The definition of Markov chains that we have seen in earlier chapters was restricted to a discrete (countable) state space, whereas in the following, continuous state spaces are allowed as well. Eventually, we want to be able to apply MCMC algorithms to the parameter space of continuous probability distributions, which makes this extension to a general state space inevitably. Throughout the rest of this chapter let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a measurable space and probability measures will always refer to this space. Of course, when dealing with real world data the most important case is $\mathcal{X} = \mathbb{R}^d$ with the euclidean topology, where the Borel sigma algebra contains all "classical" measurable sets.

3.1.1 Transition Kernels and Markov Chains

Since singletons in $\mathbb{R}^d$ have measure zero with respect to the Lebesgue measure, assigning a positive probability to a transition into a single state will be problematic, because there would not be any chance to transition into states outside of a countable subset of $\mathbb{R}^d$. We therefore exchange the notion of a transition matrix with the one of a transition kernel.

Definition 3.1 (transition kernel).
A transition kernel or Markov kernel is a function
$K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ *such that*

(i) $K(x, \cdot)$ *is a probability measure for all* $x \in \mathcal{X}$*,*

(ii) $K(\cdot, A)$ *is a measurable function for all* $A \in \mathcal{B}(\mathcal{X})$.

Due to property (i) we can formulate the probability of transitioning from any state x into some subset $A \subset \mathcal{X}$ as

$$P(X \in A|\, x) = \int_A K(x, dy) =: K(x, A).$$

Here, the integral is taken with respect to the probability measure $K(x, \cdot)$, noted as $K(x, dy)$. This more compact notation is preferred over its equivalent $dK(x, \cdot)y$. The random variable X is distributed according to $K(x, \cdot)$. Note, that for a discrete space $\mathcal{X}$ the kernel simply becomes a matrix with entries $p_{ij} = K(i, \{j\})$ which are then allowed to be positive valued.

When $K(x, \cdot)$ is an absolute continuous measure on $\mathbb{R}^n$ (i.e. absolutely continuous wrt. the Lebesgue measure), the hereby existing Radom Nikodym density is denoted as $K(x, y)$ and we can calculate the transition probability as

$$P(X \in A|\, x) = \int_A K(x, dy) = \int_A K(x, y)\, dy.$$

The n-times transition kernel, describing the probability of transitioning from any state x to a set $A \subset \mathcal{X}$ in n steps, is inductively defined by

$$K^n(x, A) = \int_{\mathcal{X}} K^{n-1}(y, A)\, K(x, dy).$$

We obtain an equivalent description on the level of probability measures, i.e. when $K(x, \cdot)$ is the object of interest:

$$K^n(x, dy) = K^{n-1}(\hat{y}, dy)\, K(x, d\hat{y})$$

As a first result, we can now proof the Chapman-Kolmogorov equation, which is essential for all upcoming calculations.

Theorem 3.1 (Chapman-Kolmogorov equation).
For all $n, m \in \mathbb{N}_0$, $x \in \mathcal{X}$ *and* $A \in \mathcal{B}(\mathcal{X})$

$$K^{m+n}(x, A) = \int_{\mathcal{X}} K^n(y, A)\, K^m(x, dy)$$

Proof. We calculate

$$\begin{aligned} K^{m+n}(x,A) &= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} K^n(y_m,A)\, K(y_{m-1},dy_m) \cdots K(x,dy_1) \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} K^n(y_m,A)\, K^{m-1}(y_1,dy_m)\, K(x,dy_1) \\ &= \int_{\mathcal{X}} K^n(y_m,A)\, K^m(x,dy_1) \end{aligned}$$

□

Definition 3.2 (Markov chain on a general state space).
Let $(\mathcal{X},\mathcal{B})$ be any measurable space (with possibly continuous $\mathcal{X}$), K a transition kernel and π_0 some initial distribution on $\mathcal{X}$. A stochastic process $(X_t)_{t\in\mathbb{N}_0}$ is called a (π_0,K)-Markov chain, if for all $A \in \mathcal{B}$, $t \in \mathbb{N}_0$ and $x_t \in \mathcal{X}$

(i) $P(X_0 \in A) = \pi_0(A) = \int_A d\pi_0(y)$,

(ii) $P(X_{t+1} \in A|X_1 = x_1, \ldots, X_t = x_t) = P(X_{t+1} \in A|X_t = x_t)$,

(iii) $P(X_{t+1} \in A|X_t = x_t) = \int_A K(x_t,dy)$.

Just like in the discrete version of a Markov chain, the third property connects the stochastic process to the transition kernel and since the transition probability is independent of t, we again call this a *homogeneous* Markov chain.

3.1.2 Invariance and Stationarity

Our goal is to generate approximate samples from a distribution utilizing MCMC algorithms. If we run a Markov chain for long enough, the chain should "reach" this distribution and from there on, should not change, i.e. if $X_n \sim \pi$ then $X_{n+1} \sim \pi$. To obtain this type of convergence we need the Markov chain to have certain properties. The following definition extends the discrete case, where the invariance of a measure is simply defined by $\pi = \mathbf{p}\pi$ for a transition matrix $\mathbf{p}$.

Definition 3.3 (invariance).
A probability (or σ-finite) measure π on $\mathcal{X}$ is called invariant with respect to a transition kernel K, if

$$\pi(A) = \int_{\mathcal{X}} K(x,A)\, d\pi(x) \qquad \forall A \in \mathcal{B}(\mathcal{X}),$$

or equivalently

$$\pi(dy) = \int_{\mathcal{X}} K(x, dy)\, d\pi(x).$$

The notions of invariant and stationary measures are often used interchangeably. Here, we will refer to measures as invariant and Markov chains as stationary. A (π_0, K)-Markov chain is called stationary or is said to have a stationary distribution π, if for some $t \in \mathbb{N}$ $X_t \sim \pi$ and π is invariant with respect to K. Obviously, once the chain has reached this distribution, it will stay there and is, in that sense, invariant of the time.

A first simple property is that $\pi(A) = \int_{\mathcal{X}} K^n(x, A)\, d\pi(x)$ for all $n \in \mathbb{N}$. Indeed

$$\int_{\mathcal{X}} K^n(x, A) d\pi(x) = \int_{\mathcal{X}} \int_{\mathcal{X}} K^{n-1}(y, A) K(x, dy) d\pi(x) = \int_{\mathcal{X}} K^{n-1}(y, A) d\pi(y).$$

Repeatedly substituting the transition probability yields the claim.

The existence of an invariant measure will later be given by the algorithmic construction of the transition kernel. The kernel for the Metropolis Hastings algorithm is chosen, such that it satisfies the detailed balance condition, which is a sufficient criterion. To guarantee convergence, however, we need the chain to fulfil two more properties, namely irreducibility and aperiodicity, which will be covered in a later section. First, we are going to look at how an algorithmic construction can look like.

3.2 Metropolis Hastings Algorithm

The goal of this section is not only to state but also to motivate the applicability of the Metropolis Hastings algorithm and see how the the Markov chain is constructed to obtain the desired stationary distribution. As we will see, a sufficient criterion for a transition kernel to inherit the stationary distribution π, is the so called *detailed balance condition*, which means that the kernel is *reversible*, in the sense of the following definition.

Definition 3.4.
A transition kernel K is reversible with respect to a probability distribution π, if

$$d\pi(x) K(x, dy) = d\pi(y) K(y, dx) \qquad \forall x, y \in \mathcal{X},$$

or equivalently

$$\int_A K(x, B) d\pi(x) = \int_B K(x, A) d\pi(x) \qquad \forall A, B \in \mathcal{B}.$$

The reason this condition plays a central role, is due to the following lemma.

Lemma 3.1.
If a transition kernel K is reversible with respect to a probability distribution π (i.e. it satisfies the detail balance condition), then π is invariant with respect to K.

Proof. We calculate

$$\int_{\mathcal{X}} K(x,A)d\pi(x) = \int_A K(x,\mathcal{X})d\pi(x) = \int_A d\pi(x) = \pi(A).$$

□

With this result, when constructing a MCMC algorithm, for the chain to have a stationary distribution, it is sufficient for the kernel to be reversible. The Metropolis Hastings algorithm does this in the most straight forward way. In the following we will set $\mathcal{X} = \mathbb{R}^n$, which is the most relevant case, and suffices for our purposes. So let g be any transition kernel on $\mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n)$ such that $g(x,\cdot)$ is absolutely continuous, i.e.

$$g(x,dy) = g(x,y)dy,$$

with density $g(x,y)$. This is the so called *proposal distribution*. Dependent on the current state x, it will suggest some new state y as a candidate, according to $g(x,\cdot)$. Remember, that for the algorithm to be applicable, it necessarily needs to hold, that $\pi = Cf$ for some constant $C > 0$ and some known function f. A proposed transition from state x to state y is accepted with probability

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)g(y,x)}{\pi(x)g(x,y)}\right\} = \min\left\{1, \frac{f(y)g(y,x)}{f(x)g(x,y)}\right\}.$$

The second equality holds true, because the normalizing constant cancels out in the fraction. This is precisely why the method works, even though the constant is not known. What is left to do now, is to show that the so defined acceptance probability of a proposed transition yields a reversible kernel. First we will derive the kernel that describes the transition behaviour of the Metropolis Hastings algorithm overall. The probability of a "*change of state*", i.e. accepting some proposal being in some state, is given by

$$P(change\, of\ state \mid X_t = x) = \int_{\mathcal{X}} g(x,y)\alpha(x,y)dy =: a(x).$$

The probability of changing the state into some specified subset A is given by

$$P(change\,of\ state, X_{t+1} \in A \mid X_t = x) = \int_A g(x,y)\alpha(x,y)dy.$$

Overall we obtain the transition kernel by calculating

$$\begin{aligned}
&K_{MH}(x,A) = P(X_{t+1} \in A \mid X_t = x)\\
&= P(change\,of\ state, X_{t+1} \in A \mid X_t = x) + P(no\,change\,of\ state, X_{t+1} \in A \mid X_t = x)\\
&= P(change\,of\ state, X_{t+1} \in A \mid X_t = x)\\
&\quad + P(no\,change\,of\ state \mid X_t = x)P(X_{t+1} \in A \mid no\,change\,of\ state, X_t = x)\\
&= \int_A g(x,y)\alpha(x,y)dy\ + (1 - a(x))\delta_x(A)\\
&= \int_A g(x,y)\alpha(x,y) + (1 - a(x))\delta_x(y)dy.
\end{aligned}$$

Or equivalently

$$K_{MH}(x,dy) = g(x,dy)\alpha(x,dy) + (1 - a(x))\delta_x(dy).$$

Note that conditional on “change of state”, the above kernel probability measure is absolutely continuous, since we have assumed the proposal distribution to be absolutely continuous. Therefore, for the first summand $g(x,dy)\alpha(x,dy)$ there exists a Lebesgue density as we have already written out in the last two equations above. Conditional on “no change of state”, the Dirac measure $\delta(\cdot)$ was introduced, which is not absolutely continuous by definition. Formally, by $\delta_x(y)$ we denote the Dirac *distribution*, which exists on $\mathbb{R}^n$ and precisely fulfils the last equality above, even though not being a density.

Lemma 3.2.
The Metropolis Hastings kernel K_{MH} is reversible with respect to π, i.e. satisfies

$$\pi(dx)K_{MH}(x,dy) = \pi(dy)K_{MH}(y,dx).$$

Proof. Since the equation obviously holds true if $x = y$, we only need to consider the case $x \neq y$. For the kernel, this corresponds to a change of state, so we only need to consider that respective part of the kernel. Formally, $\delta_x(dy)$ and $\delta_y(dx)$ are 0, if $x \neq y$. Therefore, the reversibility condition holds true, if and only if

$$\pi(dx)g(x,dy)\min\left\{1, \frac{\pi(dy)g(y,dx)}{\pi(dx)g(x,dy)}\right\} = \pi(dy)g(y,dx)\min\left\{1, \frac{\pi(dx)g(x,dy)}{\pi(dy)g(y,dx)}\right\}.$$

This can easily be seen to be true. The minimum is attained at 1 for precisely one side of the above equation. The other side then attains the minimum for the second term, where the denominator and the term before the minimum cancel out each other making both sides equal. □

Algorithm 1 General Metropolis Hastings Algorithm

```
Let $f$ be some function proportional to the target distribution $P$. Let $g_0$
denote the initial distribution and $g(\cdot|x)$ the proposal distribution.
Algorithm METROPOLIS HASTINGS($f$, $g_0$, $g$)
Initialisation Draw the first state from the initial distribution: $x_0 \sim g_0$.
loop over t
    Propose a candidate $x_{prop} \sim g(\cdot|x_t)$.
    Calculate the acceptance ratio:
        $$r = \frac{f(x_{prop})}{f(x_t)} * \frac{g(x_t|x_{prop})}{g(x_{prop}|x_t)}$$
    Generate a random number uniformly $u \sim \mathcal{U}[0,1]$.
    if $u \leq r$ then
        Accept the proposal and set: $x_{t+1} = x_{prop}$ .
    else
        Reject the proposal, i.e. $x_{t+1} = x_t$.
    end if
end loop
```

So far, by the way the kernel is constructed, we only know, that the stationary distribution exists. In the next section we are going to derive conditions for the kernel and the associated Markov chain to actually converge to the stationary distribution.

3.3 Markov Chain Properties

To derive convergence to stationarity, we need the chain to have certain properties, which we will introduce in the following sections.

3.3.1 Irreducibility

For Markov chains (X_t) and (Y_t) defined on two different discrete state spaces S_x and S_y, each having a stationary distribution π_x and π_y, respectively, a combined Markov chain on the union of both state spaces S_{xy}, which inherits the transition probabilities within a state space but does not allow to alternate between the different state spaces, has a stationary distribution given by $\pi_{xy} = \left(\frac{\pi_x}{|S_{xy}|}, \frac{\pi_y}{|S_{xy}|}\right)$. For a starting distribution that only has mass on S_x, the chain cannot converge to π_{xy}, because there will never be any mass shifted to S_y. This disjoint union of state spaces can be thought of as a disconnected graph, and the corresponding Markov chain is said to be *reducible*. In the discrete setting, a chain is *irreducible*, if transitioning from any state to any other state with an arbitrary number of steps, has a positive probability. In the continuous case, we need to weaken this definition to sets of non-zero measure.

Definition 3.5 (irreducibility).
A Markov chain (X_t) with transition kernel K is called ϕ-irreducible, if there exists a σ-finite measure ϕ, such that for every $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists $n \in \mathbb{N}$ such that $K^n(x, A) > 0$.

The next lemma illustrates the connection between ϕ and the stationary distribution.

Lemma 3.3.
For a ϕ-irreducible Markov chain with stationary distribution π, it holds that $\phi \ll \pi$.

Proof. Given some set $A \in \mathcal{B}$ with $\phi(A) > 0$, it follows that for every $x \in \mathcal{X}$ there exists $n \in \mathbb{N}$ and $\epsilon > 0$, such that $K^n(x, A) \geq \epsilon$. We can partition the space as

$$\mathcal{X} = \dot{\bigcup_k} \{x \in \mathcal{X} : K^{n_k}(x, A) \geq \epsilon_k\}.$$

Since $\pi(\mathcal{X}) = 1$, sigma additivity tells us that at least one of the sets, must have positive π measure. Denote this set as $B = \{x \in x : K^{n_j}(x, A) \geq \epsilon_j\}$, then $\pi(B) > 0$. Therefore it holds that

$$\pi(A) = \int_{\mathcal{X}} K^n(x, A) d\pi(x) \geq \int_{\mathcal{B}} K^{n_j}(x, A) d\pi(x) \geq \epsilon_j \pi(B) > 0$$

and the result follows. □

3.3.2 Cycles and Periods

Even for irreducible Markov chains it can happen, that the chain returns to some states only after a regular amount of time steps. This can, for example be due to some deterministic transitions between states, resulting in a pattern for the visiting interval for certain other states. It does not require deterministic transitions to create this undesired pattern, it is enough, if it necessary to pass through some collection of states, to be able to return to a given state. One can imagine that this *periodic* visiting pattern influences the long term behaviour and convergence properties of a Markov chain.

The goal in the following is to mathematically describe this *periodic* behaviour. We are first going to motivate the concept of *cycles* and *periods* for finite state space Markov chains. The transition matrix for an irreducible Markov chain, after possibly reordering the states, is given in block format

$$\mathbf{p} = \begin{pmatrix} 0 & \mathbf{p}_1 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{p}_2 & \cdots & 0 \\ & \vdots & & & \\ \mathbf{p}_d & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Note that if the chain is allowed to stay in a state (only one such state needs to exist), then d equals 1. If there are no self loops, but the chain can directly return to a state after visiting another state, then d equals 2 and so on. With these observations, we can define the *period* of a state i as

$$d(i) = g.c.d. \left\{ n \geq 1 : (\mathbf{p}^n)_{ii} > 0 \right\},$$

where *g.c.d.* is the greatest common divisor. For an irreducible chain, all states have the same period and we can assign the period to the whole of the chain. We call a chain *aperiodic* if it has period 1.

Basically, the extension of this definition to general state spaces builds on the intuition, that we just established for countable state spaces, where transitions are given in block matrix form.

Definition 3.6 (period and aperiodicity).
A ϕ-irreducible Markov chain has period $d \geq 2$, if d is the largest integer such that a ϕ-partition $N = \dot{\bigcup\limits_{i=1,\dots,d}} A_i$ exists, which means that $\phi(A_i) > 0$ for all $1 \leq i \leq d$ and $\phi(N) = 1$. Furthermore, it needs to hold, that $K(x, A_{i+1}) = 1$ for all $x \in A_i$ and all $1 \leq i \leq d-1$, as well as $K(x, A_1) = 1$ for $x \in A_d$. If $d = 1$, i.e. no such partition exists, then the chain is called aperiodic.

We are now going to introduce the concept of small sets. When proofing convergence results for Markov chains, utilizing the existence of a small set is a crucial step, since they come together with a minorozing measure that guarantees a positive return probability to the small set.

Definition 3.7 (small sets).
A set $C \in \mathcal{B}(\mathcal{X})$ is called a (ν, m)-small set (or simply small set) if there exists $m \in \mathbb{N}$ and a non-trivial measure ν such that

$$K^m(x, A) \geq \nu(A) \qquad \forall x \in C,\ A \in \mathcal{B}(\mathcal{X}). \tag{3.1}$$

It is always possible to put a factor $\epsilon > 0$, and require that

$$K^m(x, A) \geq \epsilon\nu(A) \qquad \forall x \in C,\ A \in \mathcal{B}(\mathcal{X}). \tag{3.2}$$

In both cases we refer to the inequality as a minorizing condition.

Remark Some authors use the latter minorizing condition as a definition of a small set. It turns out, that for some proofs, having a factor not equal to 1, is a little restrictive. However, being able to add the ϵ factor is useful for the coupling construction(3.3.4), where it acts as a probability of drawing the next state from the minorizing measure ν.

The definition of a small set is independent of one specific state, it is a property of the whole chain. It is not clear that a small set needs to exist. The fact that ϕ-irreducibility is sufficient for the existence, was proven by Meyn and Tweedie [4] and is an involved result, that we will refer to again in a later section. If there exists a small set, we can say, that for a Markov chain starting in C, there is a positive probability to return to C (or to visit any other set) after m time steps. It also holds that $\nu(C) > 0$, which follows together with the existence theorem.

3.3.3 Recurrence and Stability

When studying Markov chains, we can expect better convergence properties if the chain has good stability properties. Irreducibility of a chain is enough to guarantee that every set with positive measure is visited. The concept of *recurrency* extends this idea by requiring the chain to return to every set infinitely often on average.

Definition 3.8 (recurrence).
A ϕ-irreducible Markov chain (X_n) is said to be recurrent, if for all $A \in \mathcal{B}$ with $\phi(A) > 0$ it holds that $\sum\limits_{n=1}^{\infty} K^n(x, A) = \infty$ for all $x \in \mathcal{X}$.

Even more structure is obtained by the following notion of Harris recurrency, where an infinite number of visits to any set of states is guaranteed. We denote the first visit to a set of states by $\tau_A := \inf_{n \geq 1}\{X_n \in A\}$.

Definition 3.9 (Harris recurrence).
A ϕ-irreducible Markov chain (X_n) is said to be Harris recurrent, if for all $A \in \mathcal{B}$ with $\phi(A) > 0$ it holds that $P(\tau_A < \infty | X_0 = x) = 1$ for all $x \in X$.

The following characterization of Harris recurrency will be useful.

Theorem 3.2.
A ϕ-irreducible Markov chain is Harris recurrent if and only if for all $A \in \mathcal{B}$ with $\phi(A) > 0$, we have $P(X_n \in A$ infinitely often $\mid X_0 = x) = 1$ for all $x \in \mathcal{X}$.

Proof. Define $r_A = \sup_{n \geq 1}\{X_n \in A\}$. Assume that the chain doesn't return infinitely often, then $\exists \bar{y} \in \mathcal{X}$, $N \in \mathbb{N}$ such that $r_A = N$ and

$$P_{\bar{y}}(r_A = N) > 0.$$

This implies that

$$\int_A P_y(\tau_A = \infty) K^N(\bar{y}, dy) > 0.$$

Which means that there exists $y \in \mathcal{X}$, such that $P_y(\tau_A = \infty) > 0$. But this is a contradiction to the Harris recurrency of the chain. □

The following theorem shows that a recurrent chain only differs by a ϕ-null set from a Harris recurrent chain. We will only state the theorem, for a proof see Theorem 9.1.5 in [4].

Theorem 3.3.
If (X_n) is a recurrent Markov chain, then we can write $\mathcal{X} = H \cup N$ such that $\phi(H) = 1$ and (X_n) is Harris recurrent on H.

The following lemma further shows that a positive visitation probability almost guarantees a visit in finite time.

Lemma 3.4.
Given a Markov chain (X_n) on a state space $\mathcal{X}$ with stationary distribution π and suppose that for some $A \in \mathcal{X}$, it holds that $P_x(\tau_A < \infty) > 0$ for every $x \in \mathcal{X}$. Then for π-almost every $x \in \mathcal{X}$, we get $P_x(\tau_A < \infty) = 1$.

Proof. Assume the contrary, that there exists a set $B \subseteq \mathcal{X}$ with $\pi(B) > 0$ and $P_x(\tau_A < \infty) < 1 - \delta_1 \; \forall x \in B$. Since $P_x(\tau_A < \infty) > 0 \; \forall x \in \mathcal{X}$ there exists $B' \subseteq B$ with $\pi(B') > 0$ and some index $n_0 \in \mathbb{N}$ and $\delta_2 > 0$, such that $K^{n_0}(x, A) \geq \delta_2 \; \forall x \in B'$. Let $\eta_{B'} = \#\{k \geq 1 : X_{kn_0} \in B'\}$ denote the number of visits to B' as integer multiples of n_0. By construction, we then get $P(\tau_A = \infty, \eta_{B'} = r) \leq (1-\delta_2)^r$, which implies $P(\tau_A = \infty, \eta_{B'} = \infty) = 0$. Therefore, we get for all $x \in B'$

$$P_x(\tau_A = \infty, \eta_{B'} < \infty) = 1 - P_x(\tau_A = \infty, \eta_{B'} = \infty) - P_x(\tau_A < \infty) \geq \delta_1.$$

Now there exists $B'' \subseteq B'$ with $\pi(B'')0$, $l \in \mathbb{N}$ and $\delta > 0$ such that

$$P_x(\tau_A = \infty, \sup\{k \geq 1 : X_{kn_0} \in B'\} < l) \geq \delta \ \forall x \in B''.$$

From this, setting $n = l \cdot n_0$, we get that

$$P_x(\tau_A = \infty, X_{kn} \notin B' \ \forall k \geq 1) \geq \delta. \tag{3.3}$$

We will now find the contradiction in the following calculation. The idea is to use the stationarity of π to express the measure of A^C via a kernel integration. The transition index can be chosen arbitrarily. For $j \in \mathbb{N}$, we choose it to be jn, i.e. an integer multiple of n. This integral can now be bounded from below by a specific selection of j paths, which all end up in A^C but which are inflicted with constraints on multiples of n. These constraints are precisely chosen, such that we can apply 3.3 to each of the paths. The first constraint for each path forces the chain to once visit B'', which guarantees, that all the paths are disjoint, i.e. that they differ for at least one index. This is simply, because B'' and $(B')^C$ are disjoint.

$$\begin{aligned}
\pi(A^C) &= \int K^{jn}(y, A^C) d\pi(y) = \int P_y(X_{jn} \in A^C) d\pi(y) \\
&\geq \int P_y\left(\bigcup_{i=0}^{j-1} (X_{in} \in B'', X_{(i+1)n} \notin B', \ldots, X_{(j-1)n} \notin B', X_{jn} \in A^C)\right) d\pi(y) \\
&= \sum_{i=0}^{j-1} \int P_y(X_{in} \in B'', X_{(i+1)n} \notin B', \ldots, X_{(j-1)n} \notin B', X_{jn} \in A^C) d\pi(y) \\
&= \sum_{i=0}^{j-1} \int P_y(X_0 \in B'', X_n \notin B', \ldots, X_{(j-i-1)n} \notin B', X_{(j-i)n} \in A^C) d\pi(y) \\
&= \sum_{i=0}^{j-1} \int P_y(X_0 \in B'') P_{x \in B''}(X_n \notin B', \ldots, X_{(j-i-1)n} \notin B', X_{(j-i)n} \in A^C) d\pi(y) \\
&= \pi(B'') \sum_{i=0}^{j-1} P_{x \in B''}(X_n \notin B', \ldots, X_{(j-i-1)n} \notin B', X_{(j-i)n} \in A^C) \\
&\geq \pi(B'') j \delta
\end{aligned}$$

In the last inequality, for each summand we utilized inequality 3.3. The last three equalities follow from basic Markov chain properties. The last equality is due to the fact that π now acts as the starting distribution, the second last applies the product probability of a path and the third last is the homogeneity of the chain.

After all, we obtain a contradiction because the initially chosen j can be made arbitrary large, such that $1 < \pi(B'')j\delta \leq \pi(A^C)$.

□

3.3.4 Convergence of Markov Chains

To be able to talk about convergence, we need a suitable metric that defines the distance to the desired stationary distribution.

Definition 3.10 (total variation distance).
For two probability measures μ_1 and μ_2 on the same measurable space $(\mathcal{X}, \mathcal{B})$, we define the total variation distance as

$$\|\mu_1 - \mu_2\|_{TV} = \sup_{A \in \mathcal{B}} |\mu_1(A) - \mu_2(A)|.$$

We will abbreviate the norm $\| \cdot \|_{TV}$ simply by $\| \cdot \|$.

Theorem 3.4.
For an aperiodic and ϕ-irreducible Markov chain (X_t) with stationary distribution π, it holds for π-almost every $x \in \mathcal{X}$, that

$$\lim_{n \to \infty} \|K^n(x, \cdot) - \pi(\cdot)\| = 0. \tag{3.4}$$

Theorem 3.5.
If (X_t) is an aperiodic and Harris recurrent Markov chain with stationary distribution π, then for every $x \in \mathcal{X}$ it holds that

$$\lim_{n \to \infty} \|K^n(x, \cdot) - \pi(\cdot)\| = 0. \tag{3.5}$$

The proof of theorem 3.5 is, for example, discussed by Meyn and Tweedie [4]. Their proof utilizes a splitting construction that requires the existence of an accessible *atom*. In contrast, Roberts and Rosenthal [5], [6] follow a slightly different approach, also using a splitting construction, but instead of atoms they use small sets. We will follow their outline for the proof of theorem 3.4. To then be able to give a proof of theorem 3.5, only a minor adjustment in the proof is necessary.

Both of the mentioned proofs rely on the same existence result for small sets, that we will solely state. For a proof see Theorem 5.2.1 in [4].

Theorem 3.6 (Existence of small sets)**.**
Every ϕ-irreducible Markov chain contains a small set C with $\phi(C) > 0$.

The coupling construction As mentioned, one central tool that will be used is a so called coupling or splitting method. The idea is to independently start two Markov chains (X_n, Y_n) and make a prediction using their coupling time, that is the first time they meet in the same state. The following calculation demonstrates, how we can use the coupling time to create bounds on the convergence rate. Recall, that for every time step a Markov chain is described by a random variable X_n whose probability distribution is given by $P_{X_n}(\cdot)$.

$$\begin{aligned}\|P_{X_n}(\cdot) - P_{Y_n}(\cdot)\| &= \sup_{A\in\mathcal{B}} |P(X_n \in A) - P(Y_n \in A)| \\ &= \sup_{A\in\mathcal{B}} |P(X_n \in A,\, X_n = Y_n) + P(X_n \in A,\, X_n \neq Y_n) \\ &\qquad - P(Y_n \in A,\, Y_n = X_n) - P(Y_n \in A,\, Y_n \neq X_n)| \\ &= \sup_{A\in\mathcal{B}} |P(X_n \in A,\, X_n \neq Y_n) - P(Y_n \in A,\, Y_n \neq X_n)| \\ &\leq |P(X_n \neq Y_n)|\end{aligned}$$

The inequality follows from the fact, that both terms itself fulfil the inequality and so does the absolute value of the difference of them. Of course this inequality also holds for distributions from Markov chains with a fixed initial state. When the initial state is known, we can describe the probability distribution via the kernel, therefore we obtain

$$\begin{aligned}|P_{(x,y)}(X_n \neq Y_n)| &\geq \|P_{X_n|x}(\cdot) - P_{Y_n|y}(\cdot)\| \\ &= \|K^n(x,\cdot) - K^n(y,\cdot)\|.\end{aligned}$$

The coupling construction will consist of one chain (X_n) that behaves like $K^n(x,\cdot)$ and another one (Y_n) that behaves like $K^n(y,\cdot)$, while the second one is initialized such that $K^n(y, A) = \pi(A)$ for all $n \in \mathbb{N}$, making it directly applicable to the statement of the above theorem. Then, we can use the coupling inequality to bound their distance at time n by the probability that they are different, so that they have not coupled yet. The goal is then to construct the chains such that they have an early coupling time resulting in a good bound for the difference. We will see that for a Harris recurrent Markov chain we easily obtain such a bound. First, we are going to give the formal construction as in [5].

Due to the existence of a small set C, we have $K^{n_0}(x,A) \geq \epsilon\nu(A)$. The construction goes as follows.

Initialize $X_0 = x$, $Y_0 = y$.

Loop over n:

1. If $X_n = Y_n : X_{n+1} = Y_{n+1} \sim K(X_n, \cdot)$, $n = n+1$.
2. Else, if: $(X_n, Y_n) \in C \times C$:
   (a) with probability $\epsilon : X_{n+n_0} = Y_{n+n_0} \sim \nu(\cdot)$
   (b) with probability $1-\epsilon$:
   $$X_{n+n_0} \sim \frac{1}{1-\epsilon}\left(K^{n_0}(X_n,\cdot) - \epsilon\nu(\cdot)\right),$$
   $$Y_{n+n_0} \sim \frac{1}{1-\epsilon}\left(K^{n_0}(Y_n,\cdot) - \epsilon\nu(\cdot)\right).$$
   Construct $X_{n+1}, \ldots, X_{n+n_0-1}$ according to the distributions of X_n and X_{n+n_0} and the Markov kernel. Do the same for $Y_{n+1}, \ldots, Y_{n+n_0-1}$ and set $n = n + n_0$.
3. Else: $X_{n+1} \sim K(X_n, \cdot)$, $Y_{n+1} \sim K(Y_n, \cdot)$ and $n = n+1$.

The crucial part of the above construction is when option 2a is utilized. Here, the coupling takes place and from then on, together with option 1, the two chains will always be in the same state. Formally, let T^* denote the coupling time, then it holds that $X_n = Y_n$ for all $n \geq T^*$. Option 2b is necessary to guarantee that each chain is updated according to their kernel. This can easily be seen by calculating

$$X_{n+n_0} \sim \epsilon\nu(\cdot) + (1-\epsilon)\frac{1}{1-\epsilon}\left(K^{n_0}(X_n,\cdot) - \epsilon\nu(\cdot)\right) = K^{n_0}(X_n,\cdot),$$

equivalently for Y_{n+n_0}.

The following calculation shows, that using the coupling construction above, all that is left to do to proof convergence, is to bound the probability of a "late" coupling time.

$$\begin{aligned}
\|K^n(x,\cdot) - \pi(\cdot)\| &= \left\|\int_{\mathcal{X}} K^n(x,\cdot)d\pi(y) - \int_{\mathcal{X}} K^n(y,\cdot)d\pi(y)\right\| \\
&\leq \int_{\mathcal{X}} \|K^n(x,\cdot) - K^n(y,\cdot)\| d\pi(y) \\
&\leq \int_{\mathcal{X}} P_{(x,y)}(X_n \neq Y_n) d\pi(y) \\
&\leq \int_{\mathcal{X}} P_{(x,y)}(T^* > n) d\pi(y) \qquad (3.6)
\end{aligned}$$

If we can show that $\lim_{n\to\infty} P_{(x,y)}(T^* > n) = 0$, the convergence follows. To ultimately be able to proof convergence we need the following lemma which utilizes the aperiodicity of a Markov chain.

Lemma 3.5.
Let (X_n) be an aperiodic Markov chain with transition kernel K, C be a (ν, n_0)-small set and $S = \{n \geq 1 : \int_{\mathcal{X}} K^n(x,C)d\nu(x) > 0\}$. Then there exists $n^* \in \mathbb{N}$ such that $\{n^*, n^*+1, n^*+2, \ldots\} \subseteq S$.

Proof. First, $\int_{\mathcal{X}} K^n(x,C)d\nu(x) = P(X_n \in C \mid X_0 \sim \nu)$, so that the set S consists of all time steps that yield a positive probability to be in C, when the chain is initialized according to ν.
We define $T = S + n_0$, as the set of all time steps with a positive probability for the next state (X_{n+n_0} for some $n \in S$) to again be drawn from ν, when the chain is initialized according to ν. This corresponds to option 2a from the coupling construction. What we get, is, that the set T is closed under addition. If now $gcd(T) = 1$, it is known (p. 541 [7]) that there exists n^{**} such that $\{n^{**}, n^{**}+1, n^{**}+2, \ldots\} \subseteq T$. Then setting $n^* = n^{**} - n_0$, the claim follows. Indeed, assume that $d = gcd(T) > 1$, then for $0 \leq i \leq d-1$ define $A_i = \{x \in \mathcal{X} : \exists l \in \mathbb{N} : \int_{\mathcal{X}} K^{ld+i}(x,C)d\nu(x) > 0\}$, which yields a partition of the space, i.e. $\mathcal{X} = \dot{\bigcup\limits_{1\leq i\leq d-1}} A_i$, to which the Markov chain is periodic, giving a contradiction. □

We now have all tools at hand to proof the central convergence theorems.

Proof of Theorem 3.4. We know that a (ν, n_0)-small set C with $\phi(C) > 0$ exists, i.e. $K^{n_0}(x,A) \geq \nu(A)$ for all $x \in C$ and all $A \in \mathcal{B}$. We follow the coupling construction by initializing $(X_0, Y_0) = (x,y) \in \mathcal{X} \times \mathcal{X}$. Since $\phi(C) > 0$ we have $P_x(\tau_C < \infty) > 0$ and $P_y(\tau_C < \infty) > 0$, so there are indices $n_x, n_y \in \mathbb{N}$ such that $K^{n_x}(x,C) > 0$ and $K^{n_y}(y,C) > 0$. Utilizing Lemma 3.5 and choosing an index $n \geq n^*$, we also have $\int_{\mathcal{X}} K^n(x,C)d\nu(x) > 0$.

$$\begin{aligned}
K^{n_x+n_0+n}(x,C) &= \int_{\mathcal{X}}\int_{\mathcal{X}} K^n(z,C)K^{n_0}(y,dz)K^{n_x}(x,dy) \\
&\geq \int_C\int_{\mathcal{X}} K^n(z,C)K^{n_0}(y,dz)K^{n_x}(x,dy) \\
&\geq \int_C\int_{\mathcal{X}} K^n(z,C)\nu(dz)K^{n_x}(x,dy) \\
&\geq \int_C K^{n_x}(x,dy) > 0
\end{aligned}$$

Therefore $K^{n_x+n_0+n}(x,C) > 0$ and similarly $K^{n_y+n_0+n}(x,C) > 0$ for all $n \geq n^*$. Choosing $l \geq \max\{n_x, n_y\} + n + n_0$, we get

$$P_{(x,y)}(\tau_{C\times C} < \infty) \geq P_{(x,y)}(\tau_{C\times C} \leq l) \geq K^l(x,C)K^l(y,C) > 0 \qquad (3.7)$$

for all starting points $(x,y) \in \mathcal{X} \times \mathcal{X}$. Note that the second inequality holds, because both chains behave independent from another before coupling. Applying Lemma 3.4 to the joint chain, we get that

$$P_{(x,y)}((X_n, Y_n) \in C \times C \; infintely \; often) = 1 \qquad (3.8)$$

for $\pi \times \pi$-almost every $(x,y) \in \mathcal{X} \times \mathcal{X}$ and, therefore, for the coupling time T^*

$$lim_{n\to\infty} P_{(x,y)}(T^* > n) = 0 \qquad (3.9)$$

for $(\pi \times \pi)$-almost every $(x,y) \in \mathcal{X} \times \mathcal{X}$. To conclude the proof, we need 3.9 to hold for the marginal probability of x in the following sense. Let $G \subset \mathcal{X} \times \mathcal{X}$ denote the set of all starting points for which convergence holds in this theorem, then, as we saw, $(\pi \times \pi)(G) = 1$. Further, let $G_x = \{y \in \mathcal{X} : (x,y) \in G\}$ and let $\bar{G} = \{x \in \mathcal{X} : \pi(G_x) = 1\}$. Now for every $x \in \bar{G}$ equation 3.9 holds true. If we can show that $\pi(\bar{G}) = 1$, then the theorem is proven. Indeed, with the definition of the product measure, we calculate

$$0 = (\pi \times \pi)(G^C) = \int_{\mathcal{X}} \pi(G_x^C) d\pi(x) = \int_{\bar{G}^C} (1 - \pi(G_x)) d\pi(x).$$

Since $\bar{G}^C = \{x \in \mathcal{X} : \pi(G_x) < 1\}$ we get that $1 - \pi(G_x) > 0$ in the last integral and therefore, necessarily $\pi(\bar{G}^C) = 0$ and $\pi(\bar{G}) = 1$, proofing the theorem.

□

Proof of Theorem 3.5. Again, we know that a (ν, n_0)-small set C with $\phi(C) > 0$ exists, i.e. $K^{n_0}(x,A) \geq \nu(A)$ for all $x \in C$ and all $A \in \mathcal{B}$. We follow the coupling construction by initializing $(X_0, Y_0) = (x,y) \in \mathcal{X} \times \mathcal{X}$.

Due to the Harris recurrency it holds that $P_x(\tau_C < \infty) = 1$, $P_y(\tau_C < \infty) = 1$ for every $x, y \in \mathcal{X}$ and equivalently $P_x(X_n \in C \; infinitely \; often) = 1$ as well as $P_y(Y_n \in C \; infinitely \; often) = 1$ for every $x, y \in \mathcal{X}$.

The chains start out and behave independently from another and will do so until they couple. Each chain visits C infinitely many times. Since there exists no periodic decomposition, both chains can move freely around the space and are guaranteed to both be in C (potentially) infinitely many times so they have infinitely many opportunities to couple, i.e. $P_{(x,y)}((X_n, Y_n) \in C \times C \; infinitely \; often) = 1$ for every $(x,y) \in \mathcal{X} \times \mathcal{X}$. Formally, let n be some time, where both chains are in C, then they couple n_0 iterations later with probability ϵ, which would result in a coupling time of $T^* = n + n_0$. It follows that $\lim_{n\to\infty} P_{(x,y)}(T^* > n) = 0$ for every $(x, y) \in \mathcal{X} \times \mathcal{X}$, proofing the theorem. □

3.4 Properties of the Metropolis Hastings Kernel

Recall from the beginning of the chapter, that the transition kernel of the Metropolis Hastings algorithm is given by

$$K_{MH}(x, dy) = g(x, dy)\alpha(x, dy) + (1 - a(x))\delta_x(dy).$$

This section closes the theoretical part by proofing aperiodicity and Harris recurrence under some minor conditions for the above kernel, which will then justify the usage of this method for the experiments in the following chapter. Since our definition of aperiodicity requires the kernel to be irreducible, we are first going to show that this, indeed, holds.

Throughout this section we will assume that $\mathcal{X} = \mathbb{R}^n$, $g(x, \cdot)$ is a probability measure with Lebesgue density $g(\cdot, \cdot)$ which is positive and continuous on $\mathbb{R}^n \times \mathbb{R}^n$. Moreover, assume, that $\pi \propto f$, with

$$\pi(A) = \frac{\int_A f(x)dx}{\int_{\mathbb{R}^n} f(x)dx}.$$

Here, f is not only a measure but also denotes the density which we assume to exist w.r.t. the Lebesgue measure and to be positive over the whole space. It can be thought of as an unnormalised density of π.

Lemma 3.6.
Under the standing assumptions the Markov Chain associated with the Metropolis Hastings kernel is π-irreducible.

Proof. Let $\pi(A) > 0$, then for any $x \in \mathcal{X}$

$$\begin{aligned} K_{MH}(x, A) &= \int_A g(x, y) \min\left\{1, \frac{f(y)g(y, x)}{f(x)g(x, y)}\right\} + (1 - a(x))\delta_x(y)dy \\ &\geq \int_A g(x, y) \min\left\{1, \frac{f(y)g(y, x)}{f(x)g(x, y)}\right\} dy. \end{aligned}$$

If the minimum is attained at 1, then since $g(x, y) > 0$ we also have $K_{MH} > 0$. In the other case we get

$$K_{MH}(x, A) \geq \int_{f(y)\geq f(x)} g(y, x)dy + \int_{f(y)<f(x)} \frac{f(y)g(y, x)}{f(x)}dy.$$

Due to the positivity and continuity we have that $\inf_{y\in A} g(y, x) \geq \epsilon$ for some $\epsilon > 0$ and get

$$\begin{aligned} K_{MH}(x, A) \geq{}& \epsilon Leb(\{y \in A : f(y) \geq f(x)\}) \\ &+ \frac{\epsilon \int_{\mathcal{X}} f(y)dy}{f(x)} \pi(\{y \in A : f(y) < f(x)\}). \end{aligned}$$

Obviously, it holds that $\int_{\mathcal{X}} f(y)dy > 0$. We can also assume, that $f(x) > 0$ otherwise the minimum from the beginning is attained at 1 and the assumption follows. We now argue that not both terms in the above sum can be zero, using that π is absolutely continuous w.r.t. the Lebesgue measure.
If $Leb(\{y \in A : f(y) \geq f(x)\}) = 0$ then $\pi(\{y \in A : f(y) \geq f(x)\}) = 0$, as well, but since $\pi(A) > 0$, it holds that $\pi(\{y \in A : f(y) < f(x)\}) > 0$.
On the other hand, if $\pi(\{y \in A : f(y) < f(x)\}) = 0$, then $\pi(\{y \in A : f(y) \geq f(x)\}) > 0$ and therefore $Leb(\{y \in A : f(y) \geq f(x)\}) > 0$.

It follows that $K_{MH}(x, A) > 0$, even showing π-irreducibility for $n = 1$. □

The proof of the next lemma essentially uses the same arguments as the proof above. The central argument is, that the proposal distribution is positive on the whole space and together with a positive acceptance probability for any state that is proposed, there will never be any forced cyclic behaviour.

Lemma 3.7.
The Markov chain with respective Markov kernel K_{MH} is aperiodic.

Proof. Assume the contrary, that a periodic decomposition exists, i.e. $K(x, A_i) = 1$ for all $x \in A_{i-1}$. Using the same arguments as in the above proof, we see that

$$K(x, A_i) \geq \int_A g(x, y) \min\left\{1, \frac{f(y)g(y, x)}{f(x)g(x, y)}\right\} dy > 0,$$

which contradicts the assumption. □

To be able to show that the kernel is also Harris recurrent we need the following strong result, which gives a characterization of Harris recurrency for π-irreducible Markov chains.

Theorem 3.7.
Let (X_n) be a π-irreducible Markov chain with stationary distribution π. Then the chain is Harris recurrent, i.e.

$$\pi(A) > 0 \Rightarrow P_x(\tau_A < \infty) = 1 \quad \forall x \in \mathcal{X},$$

if and only if

$$\pi(A) = 1 \Rightarrow P_x(\tau_A < \infty) = 1 \quad \forall x \in \mathcal{X}.$$

Proof. Obviously, Harris recurrency implies the second statement. For the other direction let $\pi(A) > 0$. We assume aperiodicity, the proof only needs little adjustments for the periodic case. Let G denote the set of all starting points for which the convergence in theorem 3.4 holds, then $\pi(G) = 1$ and we get that $P_x(\tau_G < \infty) = 1 \; \forall x \in \mathcal{X}$. From theorem 3.4 we also know that

$$K^n(x, A) \to \pi(A)$$

for every $x \in G$. Since we reach the set G in finitely many steps from any starting point, the convergence above actually holds for all $x \in \mathcal{X}$. Since $\pi(A) > 0$, we now have $\sum\limits_{n=1}^{\infty} K^n(x, A) = \infty$, i.e. the chain is recurrent and due to theorem 3.3 we can write $\mathcal{X} = H \cup N$ with $\pi(H) = 1$ and (X_n) is Harris recurrent on H. We know have by assumption, that $P_x(\tau_H < \infty) = 1$, that is, the chain reaches H in finitely many steps with probability 1. Now that it's Harris recurrent on H, it also reaches A in finitely many steps with probability 1. □

Lemma 3.8.
The Markov chain associated with K_{MH} is Harris recurrent.

Proof. We already know that the chain is π-irreducible, and we can show Harris recurrency with respect to this measure. Take any $A \in \mathcal{B}$ with $\pi(A) > 0$, then $\pi(A^C) = 0$. Since $f > 0$ on $\mathcal{X}$, we also have $Leb(A^C) = 0$ and because $g(x, \cdot) \ll Leb$ for every $x \in \mathcal{X}$, also $g(x, A^C) = 0$ and $g(x, A) = 1$. Also, the probability for accepting any proposal is always positive, i.e. $\alpha(x, A) > 0$. This means that the chain will have infinitely many possibilities to move to A with a positive probability, which proofs the claim. □

3.5 Metropolis Hastings for Finding Mixing Ratios

As we have motivated in the beginning of this chapter, we want to use the Metropolis Hastings algorithm to generate samples from a posterior distribution $P(\gamma \mid D)$. Due to Bayes, we know that

$$P(\gamma \mid D) \propto P(D \mid \gamma)P(\gamma),$$

where the parameter γ corresponds to the mixing ratio. For any given mixing ratio, we have derived the calculation of the (marginal) likelihood $P(D \mid \gamma)$ in the MixedTrails section. We are free to chose any a priori distribution $P(\gamma)$ according to prior beliefs. If the prior belief is aligned with the ground truth ratio for the given data, this will speed up convergence to stationary, because the acceptance probability for samples that lie close to the truth will be higher. A non informative, i.e. a uniform prior, can always be chosen if no prior information is at hand and will also be the the choice for many of our experiments.

3.5.1 Proposal Distribution

We want to run the Metropolis Hastings algorithm, so all that is left, is to determine a proposal distribution $g(\gamma, \cdot)$. Given that the Markov chain of the Metropolis algorithm is currently in state γ_{curr}, the proposal distribution will propose some new state according to $\gamma_{prop} \sim g(\gamma_{curr}, \cdot)$. The pool of proposal distributions to choose from is restricted to those with parameter space $[0,1]^n$ and $\|\gamma\|_1 = \sum_i |\gamma_i| = 1$, such that a parameter vector drawn from this distribution can be used as a mixing ratio. A naturally good choice is the Dirichlet distribution. The support of an n-dimensional Dirichlet distribution is given by $(n-1)$ dimensional standard simplex $\Delta \subset [0,1]^n$, which precisely fulfils the before mentioned requirements.

The majority of this chapter was devoted to derive the theoretical framework and the requirements, which all the involved kernels need to satisfy for the Metropolis Hastings algorithm to converge. Due to the construction of K_{MH} the posterior $P(\gamma \mid D)$ is the invariant distribution of the Metropolis Hastings Markov chain.

Dirichlet Proposal Using a Dirichlet proposal kernel g means that $g(x, \cdot)$ is Dirichlet distributed for all $x \in \Delta$. Also, $g(x, \cdot)$ is absolutely continuous and therefore $g(\cdot, \cdot)$ is a Lebesgue density. Furthermore the density is positive and continuous throughout $\Delta \times \Delta$. This is all we needed for the convergence of the Metropolis Hastings algorithm to hold.

There are two things that we need to be aware of when using a Dirichlet distribution $Dir(\alpha)$ as a proposal. First, following the algorithm, a proposal state is drawn from a Dirichlet distribution with parameters according to the current state $\alpha \sim \gamma_{curr}$. The current state always corresponds to some mixing ratio, where all entries are smaller than 1. If the parameters of a Dirichlet distribution are smaller than 1, its density has spikes in the corners of the simplex it is defined upon, which will result in proposals that tend to have more sparse entries. Therefore, even though the current state is integrated into the proposal distribution, we don't get proposals that lie close to our current state, which is actually the core idea of this procedure. To fix this, we use a concentration factor $c \in \mathbb{R}_{>0}$ to upscale the current state, such that all entries are ≥ 1. The larger we choose this concentration factor, the more of the probability mass of $Dir(c * \gamma_{curr})$ concentrates around the current state. This concentration factor simply becomes another hyperparameter for the Metropolis Hastings algorithm that can be adjusted for different scenarios.

Secondly, a proposed state γ_{prop}, drawn from a Dirichlet distribution, can have zero entries. However, the parameters of the Dirichlet distribution need to be strictly positive. For $Dir(c*\gamma_{prop})$ to be well defined, we need to modify the zero entries. This is done by simply substracting a small fraction (0.01) from the largest entry and adding this to each zero entry of γ_{prop}.

Algorithm 2 Metropolis Hastings Algorithm with Dirichlet Proposal

```
Let $\hat{\gamma} \in \mathbb{R}^n_+$ be any positive valued vector, $c \in [1, \infty)$ some concentration
factor and a Prior probability $Pr : \Delta \to [0, 1]$.
**Algorithm** METROPOLIS HASTINGS($Pr$, $\hat{\gamma}$, $c$)
**Initialisation** Draw the first state $\gamma_0 \sim Dir(\hat{\gamma})$.
**loop over t**
    *Propose* a candidate $\gamma_{prop}$ drawn from $g_t \sim Dir(\gamma_t * c)$
    and also let $g_{prop} \sim Dir(\gamma_{prop} * c)$.
    Calculate the acceptance ratio:
        $$r = \frac{P(D|\gamma_{prop})}{P(D|\gamma_t)} * \frac{Pr(\gamma_{prop})}{Pr(\gamma_t)} * \frac{g_t(\gamma_{prop})}{g_{prop}(\gamma_t)}$$
    Generate a random number uniformly $u \sim \mathcal{U}[0, 1]$.
    **if** $u \leq r$ **then**
        *Accept* the proposal and set: $\gamma_{t+1} = \gamma_{prop}$ .
    **else**
        *Reject* the proposal, i.e. $\gamma_{t+1} = \gamma_t$.
    **end if**
**end loop**
```

Exponential Proposal Both, HypTrails and MixedTrails employed their analysis of hypotheses and their evidences over a whole range of concentration parameters, denoted as κ. As discussed, this parameter can be seen as the strength of belief in a hypothesis and it is not certain, how strong this belief needs to be, i.e. how large the κ parameter needs to be chosen to obtain comparability among different hypotheses. For that reason, it would be desirable if this parameter could be inferred, as well. We shall see in the next chapter, that the inference procedure works well for some data sets. Nevertheless, analysing the data set for different hypotheses over a range of concentration parameters prior to the inference procedure is preferable to avoid unpredictable results.

An Exponential proposal kernel fulfils all requirements for the Metropolis Hastings algorithm to converge. That is, the exponentially distributed proposal function $g(x, \cdot)$ is absolutely continuous for all $x \in \mathbb{R}_{>0}$ and yields a Lebesgue density $g(\cdot, \cdot)$ which is positive and continuous over $\mathbb{R}_{>0} \times \mathbb{R}_{>0}$. We will use this distribution to propose a new state according to the current state as follows. We draw our proposal κ_{prop} from an Exponential distribution with the current Kappa value as expectancy, i.e. from $Exp(1/\kappa_{curr})$.

In the sampling process, we use this Exponential proposal density and the Dirichlet proposal to sample the mixing ratios, simultaneously. Since both quantities are sampled independently from another, we can simply extend the calculation of the evidence to the product space with corresponding product probabilities.

When it comes to choice of a prior for the concentration parameter κ, using an uninformative prior is rather difficult, because a uniform distribution on $\mathbb{R}_{>0}$ only exists in the form of closed intervals $[a, b]$. Even though a can be chosen arbitrarily close to 0, choosing any finite bound b on the right side of the interval will always yield some prior information, since values greater than this bound will simply have probability zero. Therefore, we also use an Exponential distribution as a prior. To do this, the passed parameter, i.e. the expected value for this prior, is chosen according to the size of the data set at hand. As we are going to see in the next chapter, a "good" choice of belief value κ, is given by the number of overall transitions in the data set.

Overall, the Metropolis Hastings algorithm is adapted in a straight forward way, in particular the acceptance ratio is now calculated as follows

$$r = \frac{P(D|\gamma_{prop}, \kappa_{prop})}{P(D|\gamma_t, \kappa_t)} * \frac{Pr(\gamma_{prop})Pr'(\kappa_{prop})}{Pr(\gamma_t)Pr'(\kappa_t)} * \frac{g_t(\gamma_{prop})g'_t(\kappa_{prop})}{g_{prop}(\gamma_t)g'_{prop}(\kappa_t)},$$

which is simply an extension to the product space, where Pr' and g' denote the Exponential prior and proposal, respectively.

4

Experiments

The goal of this chapter is to apply the Metropolis Hastings algorithm to different datasets and to find the optimal mixing ratio of hypotheses. We will start by introducing a synthetic dataset which essentially can be found in the MixedTrails paper [2]. This is a synthetic dataset, constructed by random walkers on a graph. The majority of our experiments will be implemented for data sets of this kind, though they will vary in their composition, according to to different scenarios that will be covered. Before we actually run the Metropolis Hastings algorithm, we will thoroughly analyse the dataset by evaluating a selection of fixed compositions of hypotheses over a range of belief degrees. After we demonstrate our method for a variety of synthetic data scenarios and hypothesis constellations, we apply our method to the empirical Wikispeedia data set in the last section.

4.1 Synthetic Data

To generate a synthetic dataset, we are going to use random walks on a graph. There are different generative procedures to obtain graphs, for example the $G(n, m)$ model by Erdos and Reni, where a graph with n nodes is generated by placing a total of m edges uniformly at random. This procedure yields a different network topology every time the network is generated. A different approach is taken by the $G(n, p)$ model, which puts an edge with probability p

between any two of the n nodes. It can be shown, that the degree distribution $P(k) = N_k/n$ ($N_k =$ the number of states with degree k) for large networks of this kind, can either be approximated by a normal distribution or (for small p) by a Poisson distribution. These approaches are, therefore, kind of restricted. It is also known that a lot of real world networks are so called "scale free" networks, that is, their respective degree distribution follows a power law, i.e. $P(k) = k^{-\gamma}$ for some positive γ, such that the network characteristics are independent of the size of the network. The Barabasi Albert model fulfils this requirement and is, thereby, suitable for our purposes. It constructs a graph with n nodes, by successively adding m new edges for every new node. Starting with a star graph with $m + 1$ nodes, in each iteration a new node is added and connected to m other nodes, preferably to those with a high node degree. Let d_i denote the degree of node i in the current iteration and let n_{curr} be the number of nodes in the current iteration. The probability for a new node to connect to node i is then precisely given by $d_i / \sum_j^{n_{curr}} d_j$. To be able to have different types of walkers and also different hypotheses for the network, we assign three different colors - red, blue and green - to the nodes of the graph. This is done uniformly with respect to the node degree, by looping through the set of colors when adding new nodes to the graph.

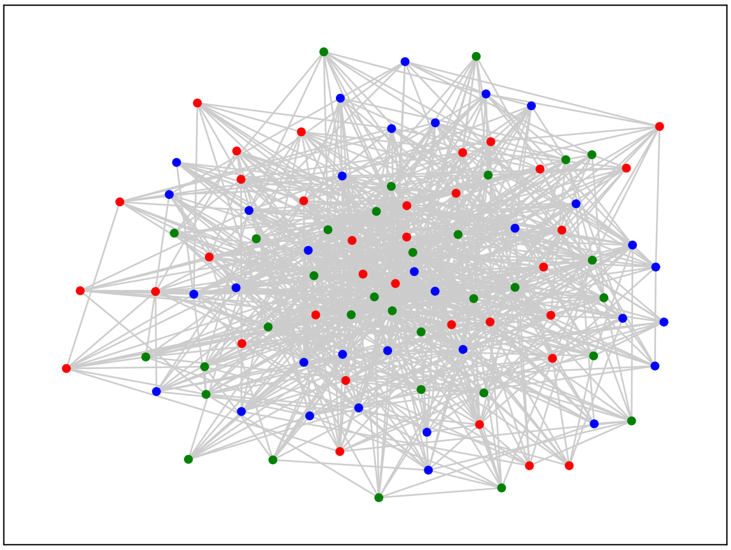


Figure 4.1: Barabasi Albert Graph with $n = 100$ and $m = 10$

Coloured Random Walks To obtain transitions, that can be used as data, we will use different coloured random walkers on the above constructed graph. We are going to consider three different types of walkers, red, blue and green. A red walker, for example, is initialized at some node of the graph and chooses the next state with equal probability between all of its red neighbours and will, therefore, always transition between red states (except for its initialised state). The blue and green walker do the same thing, simply using their own respective colour to choose the next state from. Each of these coloured walkers can be expressed via a transition matrix of size 100×100. Due to the equal distribution of the node colours, the three sub networks consisting only of nodes of the same colour, can be assumed to be topologically identical.

Hypothesis Formulation and Evaluation We are going to formulate mixed hypotheses as introduced in the last chapter, according to MixedTrails, and evaluate evidences as derived from the Mixed Transition Markov Chain model. To explain synthetic data obtained from a collection of different coloured random walkers, we are going to define hypotheses that express their belief similar to the transition probabilities of the walkers. More precisely, the transition matrix of each of the walkers is used as the belief matrix ϕ_g for one group g of the mixed hypothesis. For our scenario with 3 differently coloured walkers, we then get 3 groups, i.e. $|G| = 3$. For example, if $\mathbf{p}_r$ is the transition matrix associated with the red random walker, we set $\phi_{g_r} = \mathbf{p}_r$. Now g_r is the group that yields the belief to explain the data with red transitions only. We call this a red hypothesis and similarly define blue and green hypotheses.

To stress the difference between the overall mixed hypothesis and the hypotheses, that are encoded in each group of this mixed hypothesis, we will sometimes use the notion of a single hypothesis. Essentially, a single hypothesis refers to the belief matrix in one specific group. In that sense, the red hypothesis is a single hypothesis. A mixed hypothesis is obtained by defining group assignment probabilities for all transitions over the set of all underlying single hypotheses, which are in our case given by the red, blue and green hypothesis.

In our approach, for every observed transition $t \in D$, we will use the same group assignment probability γ_t. Therefore, our announced goal, to find the optimal mixing ratio γ for a given set of (single) hypotheses, will always be reached without differentiating between transitions. In this sense, our approach assumes, that observed trails always originate from the same transition behaviour and that this transition behaviour does not change within a trail.

4.2 Evidences for Selected Mixed Hypotheses

In the following, we are going to analyse a synthetic data set for different compositions of hypotheses. As motivated in the last Chapter, we will use the evidence as a relative scale of how good a mixed hypothesis explains the data. The dataset, that we will look at in this section, consists of 60 red walkers, 30 blue walkers and 10 green walkers each taking a total of 10 steps, which adds up to 1.000 transitions in total.

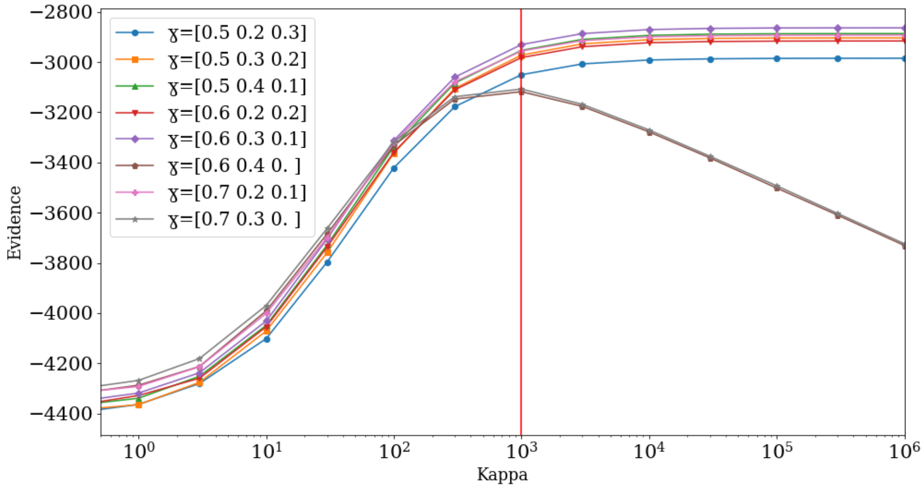


Figure 4.2: Evidence curves for a selection of mixing ratios evaluated over a range of different Kappa values indicating the degree of belief.

Figure 4.2 shows how the evidence, for a selection of mixed hypotheses, changes for different Kappa values. Here, we have chosen mixing ratios, that we expect to lie close to the ground truth mixing ratio, which is, for this dataset, given by $\gamma = [0.6, 0.3, 0.1]$. The selected mixing ratios lie close to the ground truth, in the sense that they only differ by at most 0.1 from each entry of our ground truth ratio. Plotting other mixing ratios is of course possible, but this will not give us any more insight into the data. Looking at figure 4.2, we are going to point out a few observations, that can be made already. First, we can see that for a large enough Kappa, which can approximately be located at 10^3 - where the red line is drawn - the ground truth ratio, visualized in purple, has the highest evidence. This is obviously in line with what we would expect, since this a precisely the ratio, that the data has been generated with. Also, the overall trend of the curves is that the evidence increases as the Kappa value increases. This means, that a stronger belief in any hypothesis yields a better explanation of the dataset. The only exception is given by the brown and the grey curve, which correspond to ratios $\gamma = [0.7, 0.3, 0]$ and $\gamma = [0.6, 0.4, 0]$. These two mixing ratios have a zero entry in their last argument. Speaking in terms of hypotheses, this means, that they assume there are no "green transitions", i.e. transitions into green states. But we know that actually 10% of all transitions are generated by a green walker. Now these 10% of transitions cannot be explained at all by those two hypotheses. The fact, that the hypotheses fail to explain some amount of transitions in the dataset, results in a decrease of evidence as the belief in the respective hypothesis grows. This explains, that both curves are dropping for an increasing Kappa. For all of the other mixing ratios, that do not have a zero entry anywhere and are, therefore, in principle able to explain all transitions, we can observe a different behaviour. All of those curves reach a "stationary" state, where the evidence does not really change any more, if Kappa is chosen large enough.

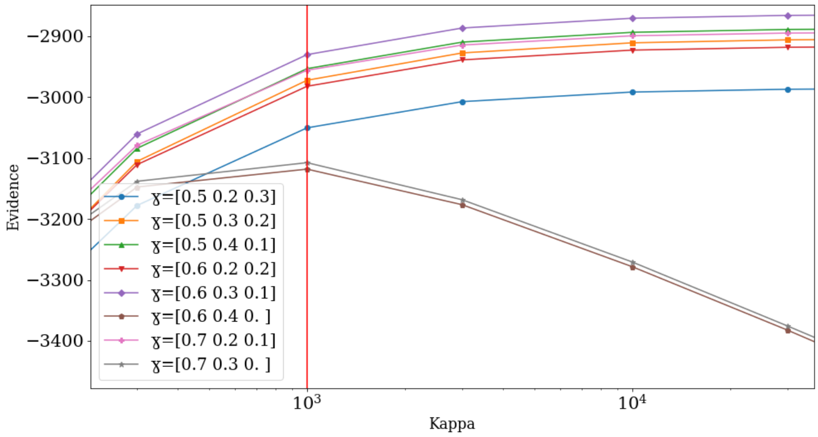


Figure 4.3: Evidences for large Kappa values

If we only look at this region of stationarity, i.e. Kappa larger than 10^3, the order of the curves seems to reflect the true ranking of the hypotheses. That is, the purple curve as the ground truth hypothesis having the highest evidence, followed by the pink and green curve corresponding to ratios $\gamma = [0.7, 0.2, 0.1]$ and $\gamma = [0.5, 0.4, 0.1]$, followed by orange and red associated with ratios $\gamma = [0.5, 0.3, 0.2]$ and $\gamma = [0.6, 0.2, 0.2]$, respectively. The blue curve, which assumes that 30% of the data consists of green transitions, has a significant lower evidence, than all of the above. Loosely speaking, the further the entries are away from the ground truth, the lower the evidence is. More precisely, the green and pink curve are pretty much having the same evidence. This is the case, because they both fail to explain the same amount of transitions, as they both differ by 0.1 from the ground truth in the first two entries.

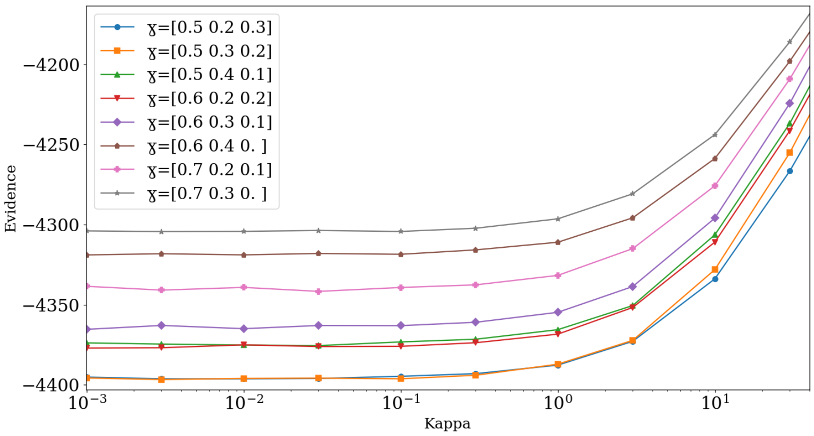


Figure 4.4: Evidences for small Kappa values

We know shift our analysis of the curves to the region of Kappa values that are smaller than 10^3. We can see, that from here on, for decreasing Kappa, the arrangement of the curves changes. While the ground truth curve drops in evidence, the grey and brown curve become those with highest evidence (see figure 4.4). The order of curves that establishes for small Kappa values is completely different from the one for large values. We find, that those mixing ratios, that concentrate their probabilities on fewer entries, yield a higher evidence and those, that distribute their probabilities more equal, yield a lower evidence. For example, $\gamma = [0.5, 0.3, 0.2]$ has a lower evidence than $\gamma = [0.7, 0.2, 0.1]$. Remember, that, as Kappa goes to 0, the hypothesis becomes more and more independent from our prior belief. Now, the only way that a hypothesis influences the evidence for small Kappa values, is via the amount of possible instantiations, that can be attained and the number of assigned data points to the corresponding groups. If there are fewer possible instantiations, then that corresponds to less uncertainty. Therefore, a comparison of mixing ratios for small Kappa values, is, in essence, a comparison of the amount of uncertainty that each mixing ratio inhibits. Those mixing ratios that have more concentrated entries, such as $\gamma = [0.7, 0.3, 0.]$, have a lower uncertainty then those with more equally distributed entries, for example $\gamma = [0.5, 0.3, 0.2]$. In other words, the evidence calculation penalizes a high uncertainty, which explains the order of the curves for small Kappa values.

Kappa and the size of the data set We are more interested in the comparison of evidences, that lie to the right side of the red line, where Kappa is greater than 10^3, because here, as discussed above, the arrangement of the curves reflects the true order. It is, therefore, important to quantify, where we can expect this boundary to lie. We know that the posterior distribution for each state i and some fixed instantiation ω of groups g, is given by $Dir(n_{s_i|\omega,g} + \alpha_{s_i|\omega,g})$. In more detail, the evidence formula for a certain group instantiation and a single group is given by

$$\prod_i \frac{B(n_{i|g,\omega} + \alpha_{i|g,\omega})}{B(\alpha_{i|g,\omega})}.$$

From this formula it becomes apparent, that the number of observed transitions n_i and the number of pseudocounts α_i are directly linked to one another. That is why we can expect some kind of proportionality between the overall number of transitions and the overall number of pseudocounts. Our example dataset from above consists of 1.000 transitions, which precisely gives a 1 : 1 dependency between these two quantities, since we have located the Kappa boundary at 10^3.

4.3 Noisy Transitions

We are now going to perturb the data set that we have used so far, by exchanging some amount of the transitions with "noisy" transitions. To do this, we are going to consider two different types of noisy walkers. The first type, we will call link walker. This walker obeys the underlying structure of the graph, by following any of the existing links in each step. It does not, however, differentiate between the colours of the nodes, which is how it differs from our coloured walkers. Since every node still has an assigned colour, a transition from this walker will automatically be either red, blue or green, even though it does not choose to be any of it. Therefore, a transition can always be assigned to either the red, blue or green hypothesis. In that sense, the noise generated by a link walker can be considered as explainable noise. We construct a dataset with 20% link noise as follows. We use 80 walkers from which 60% are red 30% are blue and 10% are green and 20 link walkers. So that 80% of the data is best explained by a mixing ratio of $\gamma = [0.6, 0.3, 0.1]$ and the other 20% being link noise. We can see from figure 4.5, that for a Kappa larger than 10^3, the purple curve no longer has a distinctively higher evidence than the others. The orange curve, corresponding to a mixing ratio $\gamma = [0.5, 0.3, 0.2]$ has just as high of an evidence as the purple one. Considering the 20% noise, we can argue, that, due to equally distributed node colours throughout the graph, these 20% of link transitions are best explained by a mixing ratio of $\gamma = [1/3, 1/3, 1/3]$. This is why the best mixing ratio for the overall data set is no longer at $\gamma = [0.6, 0.3, 0.1]$, but is slightly shifted towards $\gamma = [1/3, 1/3, 1/3]$, yielding $\gamma = [0.5, 0.3, 0.2]$ as the ratio of highest evidence.

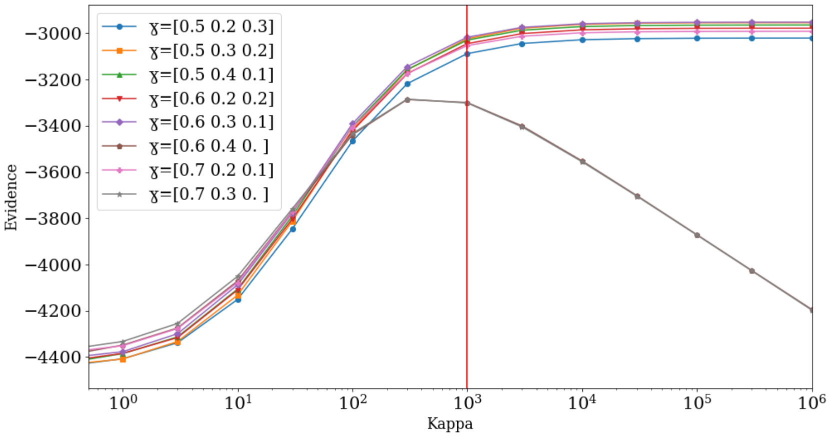


Figure 4.5: Evidence curves with 20% link noise

As a second example of how to integrate noise into the data set, we will now consider a different type of walker, one that does not follow any structural properties of the underlying graph. The simplest way to do this, is, to allow to transition between any two states from the graph, even if the nodes are not connected. We will refer to transitions of this type as teleport noise. Figure 4.6 shows how the evidence curves behave, when, as before, substituting 20% of the base transitions with teleport noise. Again, the base transitions follow a $[0.6, 0.3, 0.1]$ split, referring to the proportion of red, blue and green walkers, respectively. We point out two major differences, that become apparent when comparing the two scenarios. First, we see that all curves are decreasing in evidence for Kappa values that are large enough. The first Kappa value, for which all the curves are decreasing, again, is roughly at 10^3, at the point where the number of observed transitions and the pseudocounts are in balance. The decrease in evidence is due to the fact, that 20% of the transitions, precisely those that are generated by the teleport walker, cannot be explained by our mixed hypothesis, because an observed transition between two non neighbouring states, can neither be explained by a red, blue or green hypothesis. Due to the lack of explainability, increasing the belief into the hypothesis (i.e. increasing Kappa) for any mixing ratio results in a decrease in evidence. One thing that is different now, compared to the dataset perturbed with link noise, is, that now, the purple curve, corresponding to the ground truth ratio ($\gamma = [0.6, 0.3, 0.1]$) of the non noisy transitions, has the highest evidence for Kappa larger than 10^3. This is to be expected, because as mentioned, the teleport noise cannot be explained by our mixed hypothesis, and there exists no other ratio (that explains the teleport noise), that the ground truth ratio could be shifted towards to, to obtain a better overall explainability.

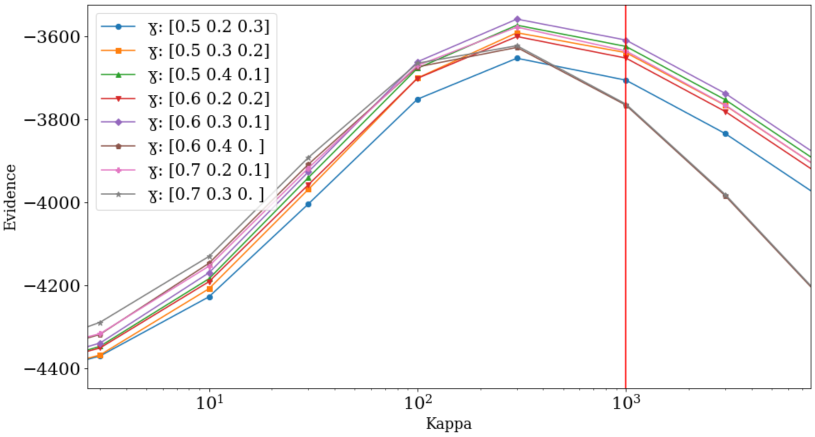


Figure 4.6: Evidence curves with 20% teleport noise

4.4 Finding Optimal Mixing Ratios

So far, we have analysed the data set and saw how the evidence curves behave for different mixing ratios and different Kappa values. We also showed, that there exists some proportionality between the size of the data set and the Kappa threshold, for which we can restore the order of the evidence curves, that we would expect. In this section, we want to use the Metropolis Hastings algorithm to automatically find the optimal mixing ratio, i.e. the ratio that yields the highest evidence. As a first approach, we fix $\kappa = 10^3$. For this fixed Kappa, figure 4.7 shows the evidences of all possible mixing ratios discretized with a step size of 0.1. For general mixing ratios, let p denote the first entry of γ, q the second entry and the third entry is then given by $1-p-q$, which gives $\gamma = [p, q, 1-p-q]$. Figure 4.7 gives us even more insight to the dataset. Firstly, we see, that the highest evidence is attained for $\gamma = [0.6, 0.3, 0.1]$, which obviously is what we would expect. The further the ratios lie away from this maximum, the lower the evidence becomes. The bottom row corresponds to mixing ratios, where the first entry p is set to 0, i.e. all the hypotheses that cannot explain red transitions. We can see, that here, the evidence takes the lowest values. This makes sense, because the majority (60%) of the data set consists of red transitions. In the bottom left corner, where both p and q are zero, the evidence attains its minimum. The mixing ratio is given by $\gamma = [0, 0, 1]$, which means that only the green transitions are covered and since they only make up 10% of the data, it is reasonable that the minimum is attained for this mixing ratio. The column on the very left consists of all mixing ratios with $q = 0$. The evidences are also pretty low, because the blue transitions (30% of the data set) are not explained by these hypotheses. Lastly, the diagonal consists of hypotheses that do not explain the green transitions, which, again, contribute to only 10% of to the data. Therefore, the values are not as low as for the other two extreme cases, where one colour is left out in the mixed hypothesis.

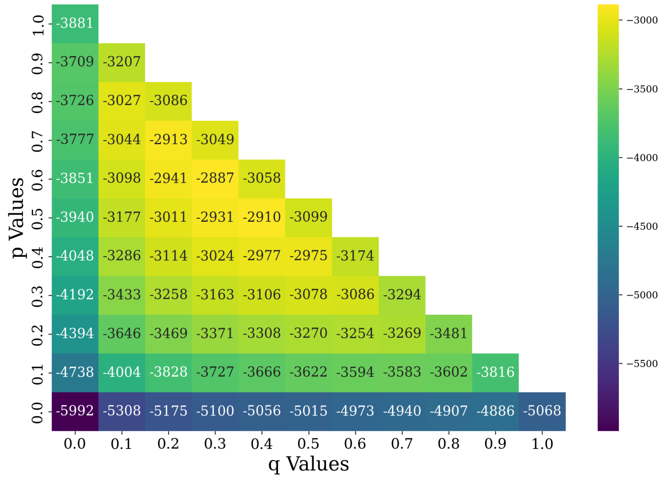


Figure 4.7: Heatmap of Evidences for Kappa = 10^3

We now want to find the optimal mixing ratio automatically using the Metropolis Hastings algorithm. To run the algorithm, we choose a uniform prior $P(\gamma) \sim Dir(1, 1, 1)$ and initialize the first state by also drawing uniformly at random, i.e. γ_0 drawn from $Dir(1, 1, 1)$. We use a Dirichlet proposal as discussed in chapter 3. Recall the concentration factor c that is incorporated into the proposal distribution influencing how close the proposed states lie to the current state. We are going to choose this concentration factor to be $c = 100$ throughout all of our experiments. In each iteration of the algorithm, the acceptance ratio is calculated, by evaluating the likelihood $P(D \mid \gamma)$ for $\kappa = 10^3$.

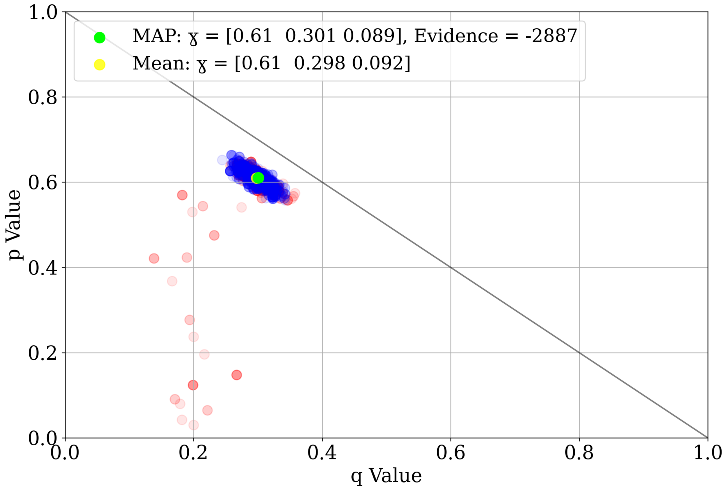
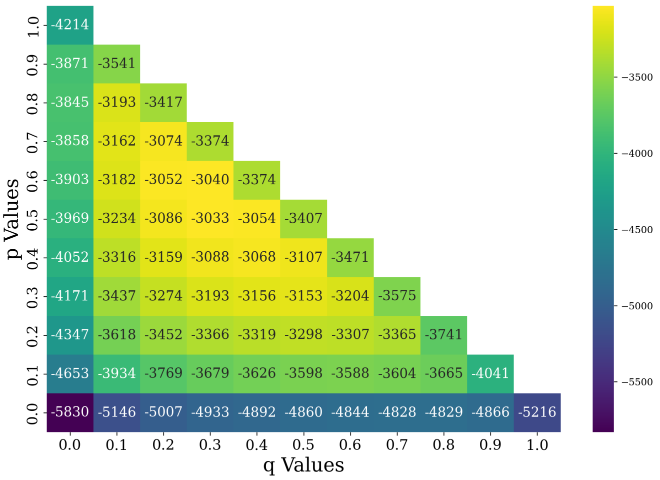


Figure 4.8: Samples generated from the Metropolis Hastings algorithm

Figure 4.8 shows all of the samples that have been accepted from the algorithm. For this particular run we used 3.000 iterations, and had a total of 550 accepted samples. As mentioned, we used a concentration factor $c = 100$ for the Dirichlet proposal distribution. The red dots are the samples, that we consider to be part of the burn in phase. For these samples we assume that the Markov chain has not yet reached stationarity and they can, therefore, be disregarded. They still give insights to the progress of reaching the stationary distribution. We can see from the figure, that the algorithm starts at around $\gamma = [0, 0.2, 0.8]$ and then makes its way closer and closer to the optimal state $\gamma = [0.6, 0.3, 0.1]$. For this run, we consider roughly the first 40% of iterations to belong to the burn in phase. The blue dots are those samples that we assume to be sampled from the posterior (i.e. the stationary distribution) and that are actually used to give further estimates. We can already see from figure 4.8, that the location and arrangement of the dots, concentrates around $\gamma = [0.6, 0.3, 0.1]$. To validate this observation, we determine two classical point estimators, the mean and the MAP (maximum a posteriori) estimator. They are visualised by the yellow and green dot, respectively. We can see, that both of the estimators predict similar results, that only differ slightly from the (theoretically) optimal solution $\gamma = [0.6, 0.3, 0.1]$. The small deviation is not significant. The MAP yields an evidence of -2887, matching the the evidence given in figure 4.7 for the optimal mixing ratio $\gamma = [0.6, 0.3, 0.1]$.

4.5 Noise Data and Kappa Influence

In the following we will apply the Metropolis Hastings algorithm to the two noisy data sets that we have introduced in section 4.3. We have already discussed, that the transitions generated from a link walker are actually covered by the explanatory scope of any mixed hypothesis, because a transition, that follows the graph structure, can always be assigned to some coloured hypothesis. This results in the mentioned shift of the mixing ratio of highest evidence away from $\gamma = [0.6, 0.3, 0.1]$. We now again fix $\kappa = 10^3$ and compare the evidences for all equidistant (0.1 distance) mixing ratios.



p \ q	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	-4214										
0.9	-3871	-3541									
0.8	-3845	-3193	-3417								
0.7	-3858	-3162	-3074	-3374							
0.6	-3903	-3182	-3052	-3040	-3374						
0.5	-3969	-3234	-3086	-3033	-3054	-3407					
0.4	-4052	-3316	-3159	-3088	-3068	-3107	-3471				
0.3	-4171	-3437	-3274	-3193	-3156	-3153	-3204	-3575			
0.2	-4347	-3618	-3452	-3366	-3319	-3298	-3307	-3365	-3741		
0.1	-4653	-3934	-3769	-3679	-3626	-3598	-3588	-3604	-3665	-4041	
0.0	-5830	-5146	-5007	-4933	-4892	-4860	-4844	-4828	-4829	-4866	-5216

Figure 4.9: Heatmap of Evidences for Kappa = 10^3 for Link Noise Data

The heatmap 4.9 illustrates the distribution of the evidence for the different ratios. The result is pretty similar to the heatmap for non noisy data (see figure 4.7) and the low evidence regions can be explained using the same arguments. The main difference is the location of the highest evidence, which is now attained for $\gamma = [0.5, 0.3, 0.2]$, and is due to the existence of link transitions.

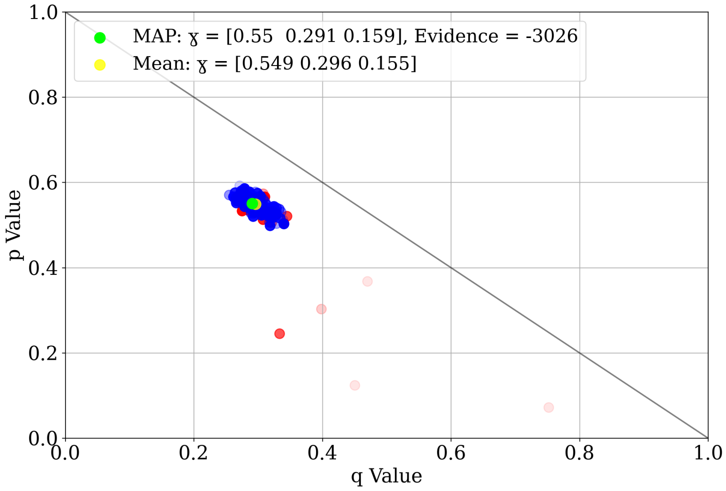


Figure 4.10: Metropolis Hastings samples on data with 20% link noise

This expected shift of the mixing ratio can also be recovered from the Metropolis algorithm. Figure 4.10 shows the progress of the samples as they get closer to stationarity. We have again used 3000 iterations for the algorithm and had a total of roundabout 520 accepted samples, of which the first 40% are considered to belong to the burn in phase. If we look at mean and MAP estimators, that are evaluated from the posterior samples (blue), we can see, that they both predict a quite similar mixing ratio. The mixing ratio predicted by the MAP is $\gamma = [0.55, 0.291, 0.159]$ and scores an evidence value of -3026, which is better than any value we can find in figure 4.9. The grid, used by the heatmap, is too coarse to cover this precise ratio and therefore misses the optimal value. This clearly is an advantage of the Metropolis Hastings algorithm, because it is not known beforehand, how close the distances between the ratios in the heatmap need to be chosen. The sampling approach is, in that sense, more flexible as it does not require any knowledge about the grid size.

4.5.1 Kappa as a Parameter

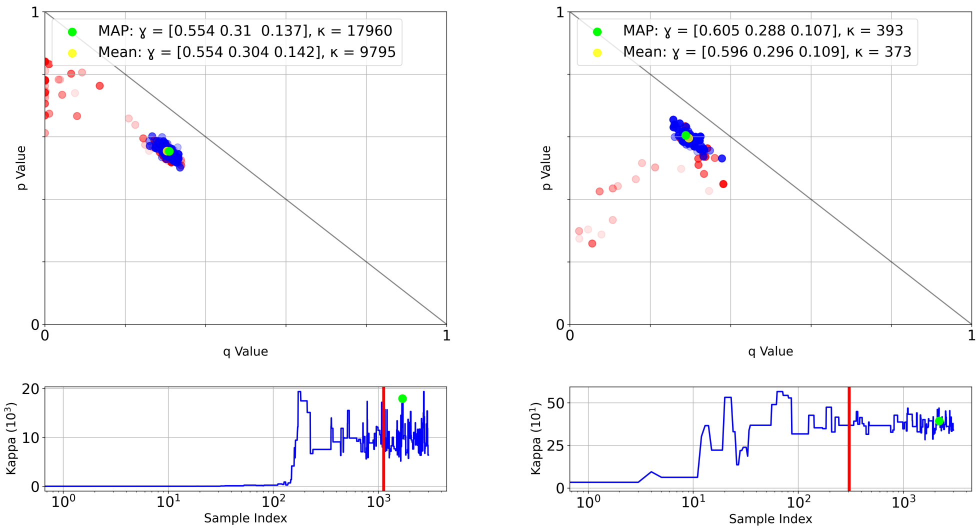
For all the above experiments and runs we have used a fixed Kappa value of 10^3. We have discussed how this is justifiable in regard to the size of the data set and how the evidence curves behave for Kappa values that are larger than this threshold. Again, looking at the evidence curves from figure 4.5 or figure 4.2, we see, that the part of the curve to the right of this threshold, actually seems to attain a global maximum over all Kappa values somewhere in this region. It therefore seems plausible to ask, if we can find the best mixing ratio and the optimal Kappa value simultaneously. We try to answer this by running the Metropolis Hastings algorithm for the two independent quantities: Mixing ratio and Kappa value. To generate proposals for the mixing ratio, we use a Dirichlet distribution just as before. To obtain proposals for the Kappa value, we use an Exponential distribution $Exp(\lambda)$. In each iteration we choose the expected value of the Exponential distribution, that generates the proposals, to be the current Kappa value, i.e. $1/\lambda = \kappa_{curr}$. As mentioned, the two quantities are sampled independently from another and the sampling space gets extended to $\Delta \times \mathbb{R}_{>0}$, where $\Delta \subset [0,1]^3$ denotes the standard simplex as the support of the Dirichlet distribution.

For the first experiment we stick with the data set, that consists of 20% link noise. Figure 4.11a shows the results of a run of the Metropolis algorithm with again 3000 iterations. For this run we used a Dirichlet proposal distribution with concentration factor $c = 100$ and a uniform prior. For the Kappa value we used an Exponential proposal distribution and an exponentially distributed prior with expected value of $1/\lambda = 1000$. This prior is chosen according to our belief, that this (Kappa = 1000) is the threshold for the given data set, for the evidences to reach their "true" arrangement.

It would also be interesting to choose a uniform prior for the Kappa sampling, but since Kappa is not naturally bounded from above, it would be difficult to determine an upper bound for a uniform distribution on the positive real line. Determining any finite upper bound for a uniform distribution, would already result a in prior that represents some kind of belief, as a Kappa value larger than this bound will never be accepted.

The number of accepted samples for this run drops to about 180, which is due to existence of a further dimension in the sampling space. The first part of figure 4.11a shows the progress of the mixing ratio samples generated by the algorithm, which, apart from a different starting point, does not differ too much from figure 4.10, that we already saw for this data set. The progress of the generated Kappa samples is visualized below. The plot shows the Kappa value for a given iteration on a logarithmic scale. Again, the red dots indicate the samples that are considered to be part of the burn in phase, accordingly chosen is the red line in the Kappa visualization below. The Kappa samples that lie to the left of that red line can be considered to belong to the burn in phase.

For this run we have initialised Kappa with a value of 10 and we can see that it does not take long for the Kappa values of the accepted samples to rise. After about 50 iterations of the algorithm, a first slight increase can be observed. It only takes about a little more than 100 iterations for the algorithm to reach a Kappa value of 10.000. This is clearly larger than the determined threshold and chosen prior expectancy of 10^3. With our discussion about the evidence curves, we can therefore assume all the samples that lie in this region to obey to the true order of evidences.



(a) Samples for mixing ratio and Kappa value obtained from a data set with noisy link transitions.

(b) Same set up, but here the data was perturbed with teleport transitions.

Figure 4.11: Sampling mixing ratios and Kappa values simultaneously.

Note also, that even though the Exponential prior has an expected value of 1.000, the Kappa values of the accepted samples are settling at around 10.000 with spikes as high as 17.000 and as low as 4.000. Looking at the evidence curves in figure 4.5, this behaviour makes sense. First, we can see that the evidence on the precise threshold of $\kappa = 10^3$ is significantly lower for every mixing ratio than for $\kappa \geq 10.000$. For values larger than 10.000 though, there does not seem to be a significant change in evidence any more.

For this run the MAP and the mean estimate give similar results, only that they now also provide a Kappa value. While the mixing ratios are relatively close to one another, the Kappa values differ a bit more. But as mentioned, both Kappa values lie in a range, where the evidence difference is not significant.

Since the evidence is not becoming significantly higher for Kappa values larger than 10.000, the exponential prior with expected value of 1.000 weighs strong enough in the calculation of acceptance probability, to bound the Kappa values ($\leq$ 20.000) of the accepted samples. However, the Kappa values are not bound to the prior expectancy (1.000), since the evidence values still significantly increase for Kappa $\geq$ 1.000.

We are now going to consider the data set that consists of 20% teleport noise, with evidence curves illustrated by figure 4.6. Recall, that the evidence curves started to decrease for an increasing Kappa larger than the threshold of 10^3. For this data set we again run the Metropolis Hastings algorithm with 3.000 iterations. The proposal and prior distributions and their parameters are chosen precisely as before for the link noise data set. In particular, we again choose an Exponential prior with expected value of $1/\lambda = 1.000$. The number of accepted samples drops even further to about 100, which is due to the fact, that proposals for Kappa values that are arbitrary large, are no longer accepted, because of the mentioned decrease in evidence. We can also see this in the plotted progress of the Kappa values in figure 4.11b. We again initialised with a Kappa value of 10. We see that it starts to increase after only a few iterations and the values settle around a value for Kappa at around 350. Spikes towards larger and smaller values still exist, but their amplitude is rather low. This behaviour is to be expected, when looking at the evidence curves of the data set (figure 4.6). We can see here, that the maximum value is attained for precisely a Kappa value of 300 (or even a bit larger, but the next Kappa the evidence is evaluated for, is 1.000). Also, there is not much room for Kappa values to deviate from this maximum, because the evidence starts to decrease instantly when moving away from this point.

Another thing to point out, is the fact, that the accepted samples accumulate around the ground truth ratio of $\gamma = [0.6, 0.3, 0.1]$. Again, this is to be expected when considering the corresponding evidence curve, which yields this precise ratio as a maximum.

4.6 Multi Modal Sampling

A great advantage of the Metropolis Hastings algorithm over any kind of gradient based optimization procedure, is, that it is guaranteed to “find” every mode that exists on the evidence surface. Of course, this guarantee is of theoretical nature and we still need to adjust hyperparameters as well as prior and proposal distribution properly to generate those samples in a foreseeable amount of time. In this section, we want to demonstrate, that the algorithm is indeed capable of finding distinctive modes for a data set and corresponding hypotheses.

4.6.1 Multi Modal Evidence Surface

Finding the highest evidence is essentially an optimization problem over the space of mixing ratios where the evidence corresponds to a negative loss function that needs to be maximized. The shape of the evidence surface results from the data set, as well as the underlying single hypotheses that are used by the mixing ratios. To construct a multi modal evidence surface, we construct a synthetic data set with the same set up as in the experiments done earlier. We again use a total of 100 walkers each taking 10 steps on a Barabasi Albert graph with 100 nodes and equally distributed node colors, red, blue and green. This time, the amount of coloured walkers is balanced, so that 1/3 of the total number of walkers is assigned to each colour.

The single hypotheses that are used now, are different from the ones we have used earlier. We are going to consider a total of 4 different single hypotheses, yielding a loss surface that could only be visualized in 4 dimensions. Of course, the convergence of the algorithm is independent of the dimension of the space and it is still able to generate samples from the posterior even if the space is high dimensional. The price for sampling in a high dimensional space is a longer run time of the algorithm until it reaches stationarity, because more samples will rejected.

As mentioned, for a mixture of 4 hypotheses we are not able to visualize the loss surface but we are still able to visualize the samples that we obtain from the Metropolis Hastings algorithm, as the mixing ratios are now given by $\gamma = [p, q, s, 1-p-q-s]$ - a 3 dimensional space. We now define the single hypotheses. The first one, called red-blue hypothesis, is a mixture of the red and blue hypothesis. It is obtained by assigning equal positive probabilities to transitions to neighbours that are either red or blue. More precisely, for node i, the i'th row of the underlying belief matrix ϕ has a positive probability in every entry that corresponds to either a red or blue neighbouring state. This probability is the same for every entry in that row and is obtained by normalizing over the total amount of neighbours of state s_i. The second hypothesis is called the red-green hypothesis and is obtained in a similar manner as the red-blue hypothesis, only that red and green neighbours are taken into consideration. The third and forth are simple blue and green hypotheses, respectively, as we have used earlier. We will change the notation of the mixing ratios to $\gamma = [rb, rg, b, g]$, so that the group affiliations can be retraced immediately.

Note, that in this scenario, the red transitions can only be explained in combination with blue or green transitions. Since the data set consists of equal amounts of all three transition colours, we expect the highest evidence for those mixing ratios that precisely account for an equal appearance of those three different coloured transitions. That is, a ratio of $\gamma = [\frac{2}{3}, 0, 0, \frac{1}{3}]$ as well as $\gamma = [0, \frac{2}{3}, \frac{1}{3}, 0]$ explains precisely $\frac{1}{3}$ of each transition colour. We expect these two points to have the highest possible evidence, because any ratio, that shifts away from explaining $\frac{1}{3}$ of each transition colour, will simply lack explainability proportional to the resulting offset.

Actually, every point that lies on the line drawn between these two modes will have the same evidence. A point $\hat{\gamma}$ on that line can be expressed as a convex combination of the two modes, i.e. $\hat{\gamma} = [\frac{2}{3}t, \frac{2}{3}(1-t), \frac{1}{3}(1-t), \frac{1}{3}t]$ and we can see, that for any $t \in [0, 1]$ adding the first two entries gives $\frac{2}{3}$, so that $\frac{2}{3}$ of the data is explained by either the red-blue or the red-green hypothesis, and, therefore, due to equal probability distribution within both of these hypotheses, $\frac{1}{3}$ is explained by a red hypothesis. Also $\frac{1}{2} \cdot \frac{2}{3}t + \frac{1}{3}(1-t) = \frac{1}{3}$ is the amount of blue transitions and $\frac{1}{2} \cdot \frac{2}{3}(1-t) + \frac{1}{3}t = \frac{1}{3}$ is the amount of green transitions that are explained from the data set for any given t.

For example, for $t = \frac{1}{2}$, we obtain the point that lies in the middle of that line, given by $\gamma_0 = [\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}]$.

4.6.2 Adjusting the Evidence Surface According to Prior Settings

As just discussed, the resulting evidence surface does not have distinct modes, but a ridge along which all points are maxima. Even though, ultimately, we are interested in finding maximal evidence values, the objective from an algorithmic point of view is different. The stationary distribution, that is to be approximated by the Metropolis Hastings algorithm, is a posterior, which itself is a combination of likelihood (exp(evidence)) and prior distribution:

$$P(\gamma \mid D) \propto P(D \mid \gamma)P(\gamma).$$

With this in mind, the intuition of adjusting the prior is actually to change to objective of the algorithm, which is the posterior. Samples that are generated, follow the shape of the posterior, a “loss surface”, that takes into account both, evidence and prior. For the experiments we have done so far, we always used a flat prior, i.e. a uniform distribution, for which the posterior is actually proportional to the likelihood, because $P(\gamma)$ is constant.

In contrast to adjusting the prior, changing the proposal distribution, which makes up the final component of the Metropolis Hastings algorithm, does not change the posterior and, therefore, does not influence the stationary distribution which we generate samples from.

Sparse Dirichlet Prior

For the Dirichlet distribution $Dir(\alpha)$, setting the parameters to be smaller than 1, i.e. $\alpha_i < 1$ for all i, gives a distribution, that has probability peaks in the corners of the simplex it is defined upon. Since corners correspond to unit vectors, which are sparse by definition, we can utilize such a distribution to manipulate the algorithm to favour more sparse points. For the above data set, we choose the prior to be $Dir((0.1, 0.1, 0.1, 0.1))$ distributed. By doing this, the two points $\gamma = [0, \frac{2}{3}, \frac{1}{3}, 0]$ and $\gamma = [\frac{2}{3}, 0, 0, \frac{1}{3}]$ are preferred over any point lying on the connecting line, because they are more sparse and, therefore, are closer to the modes of the posterior. Figures 4.12 and 4.13 show that, indeed, these two more sparse points are preferred in the sampling procedure.

4.6.3 Sparse Proposals

If we run the algorithm using a proposal distribution that proposes states close to the current state, it can become difficult and time consuming for the algorithm to leave a maximum once reached. Figure 4.12 shows this problem, where a Dirichlet proposal distribution with concentration factor $c = 100$ is used.

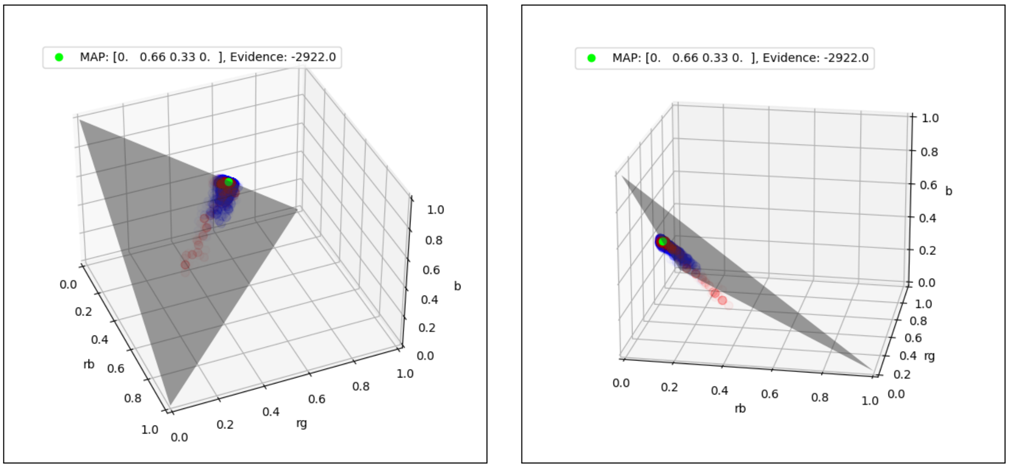


Figure 4.12: Samples obtained from the Metropolis Hastings algorithm with 10.000 iterations started at $\gamma_0 = [\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}]]$. The red dots show the accepted samples from the first ~ 600 iterations, which it takes for the algorithm to find the mode at $\gamma = [0, \frac{2}{3}, \frac{1}{3}, 0]$.

The Metropolis Hastings algorithm is initialised at the point $\gamma_0 = [\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}]$ (lying precisely in the middle of the ridge) and runs for 10.000 iterations using a Kappa value of 1000 and the mentioned sparse prior $Dir((0.1, 0.1, 0.1, 0.1))$. The axes are named with respect to the corresponding entries of the mixing ratio vector $\gamma = [rb, rg, b, g]$ (or group assignment probability vector). The grey triangle indicates the boundary of the simplex that the mixing ratios lie within. We can see, that the algorithm only needs a little more than 500 iterations, visualised by the red dots, to find the maximum located at $\gamma = [0, \frac{2}{3}, \frac{1}{3}, 0]$ as indicated by the MAP in green. Yet, it does not find the second maximum. Even though some of the accepted samples (in blue) move away from the mode, they never come far enough towards their starting point, to hop over to the second mode. Theoretically, simply running the algorithm for a long enough time, it will eventually reach the second maximum. However, this approach seems impractical under the given observations and number of iterations already used.

As an alternative we use a sparse Dirichlet distribution, which proposes states independently from the current state. With this approach, we can certainly expect a much lower acceptance rate, but it should guarantee us to find both modes in a reasonable amount of time. For this experiment we chose a $Dir((0.8, 0.8, 0.8, 0.8))$ proposal distribution in each iteration. Compared to uniform proposals, this sparse distribution should make proposals that are more aligned with the actual evidence surface and should therefore yield a higher acceptance rate than uniform proposals.

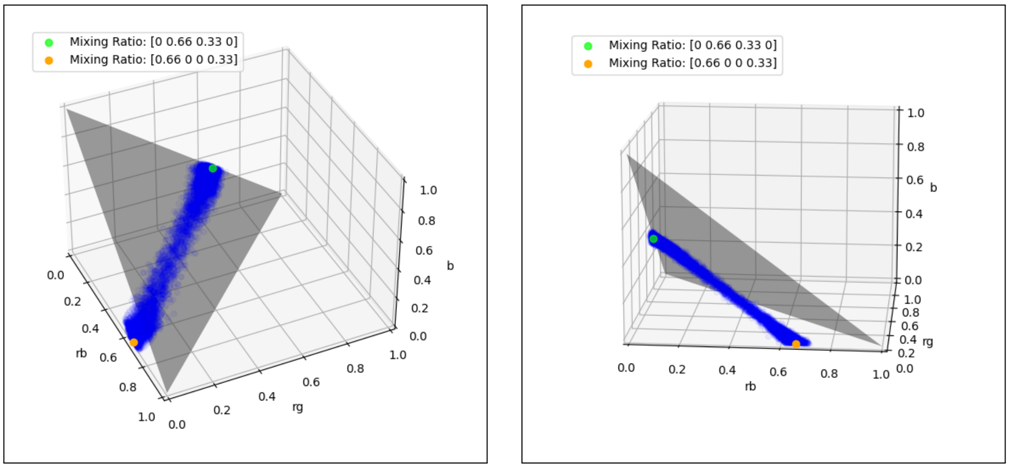


Figure 4.13: Samples obtained from the Metropolis Hastings algorithm with sparse proposal and prior distribution. The first 6.000 accepted samples are disregarded and considered to belong to the burn in phase. Visualized are the remaining ~ 6.000 accepted samples.

We carry out one final experiment with the specified set of parameters, proposal distribution and a sparse prior as before, i.e. $Dir((0.1, 0.1, 0.1, 0.1))$. We let the algorithm run for 10 million iterations. We get an acceptance rate of ~ 0.0012107, or, equivalently, a total of 12.107 accepted samples. As expected, the acceptance rate is rather low, which is due to the sparse proposal distribution. Figure 4.13 shows the results from this run, where we have disregarded the first 6.000 accepted samples (burn in) and only consider the remaining ones. The two mentioned modes $\gamma = [\frac{2}{3}, 0, 0, \frac{1}{3}]$ and $\gamma = [0, \frac{2}{3}, \frac{1}{3}, 0]$ are visualized by the green dots. We can see, that this time, samples are generated from both modes in a balanced manner. The samples are located more dense around the two modes and less dense on the connecting line. This observation coincides with our expectation, as we have motivated the theoretical preference of the algorithm towards more sparse points. Overall, with this experiment, we were able to show the applicability of our method to datasets and hypotheses, that have more than one maximum.

4.7 Empirical Data: Wikispeedia

As a final appliance of our approach we consider the Wikispeedia data set, which is a freely available data set provided by the Stanford university. This is an empirical data set obtained from users navigating through a sub network of Wikipedia with the goal to reach a specified target page with a minimal number of clicks. For each run the starting point is chosen at random. The underlying network consists of a total of 4.600 pages that follow the link structure from Wikipedia. MixedTrails [2] also carries out experiments for this data set. They confirm the hypothesis posed by authors West and Leskovic [8], that within a trail, users first navigate towards hubs and later according to semantic or textual similarity. For comparability reasons, we are going to modify the data and select certain trails in the same manner as MixedTrails. That is, we remove back clicks, but keep the corresponding forward click, which would be undone by the back click. Furthermore, only trails (click sequences) of length between 3 and 8 are considered. The resulting data set contains around 25.000 trails with an average length of 5.6 clicks and a total number of 130.000 transitions (clicks).

4.7.1 Hypothesis Conception

We construct hypotheses similar to the way in MixedTrails. The first hypothesis represents the belief, that users prefer to navigate to hubs. This belief is expressed as a group transition probability, in the matrix ϕ_{deg}. Entries are set to zero for non neighbouring states. For neighbouring states, the transition probabilities are chosen proportional to the node degree (in and outgoing edges) of each neighbour. The second hypothesis expresses the belief to prefer transitions to pages, that have a high textual similarity, encoded in ϕ_{sim}. Here, for neighbouring states, the transition probability is set proportional to the cosine similarity of the computed $tf - idf$ ("term frequency - inverse document frequency") vectors, where terms with a document frequency of more than 80% are disregarded and sublinear scaling is applied to the tf vectors. Lastly, the link hypothesis ϕ_{link} is used as a sort of reference hypothesis. In [2] it is shown, that navigating according to ϕ_{deg} for the first two clicks and according to ϕ_{sim} for all the following clicks, yields a higher evidence, than navigating according to ϕ_{link}.

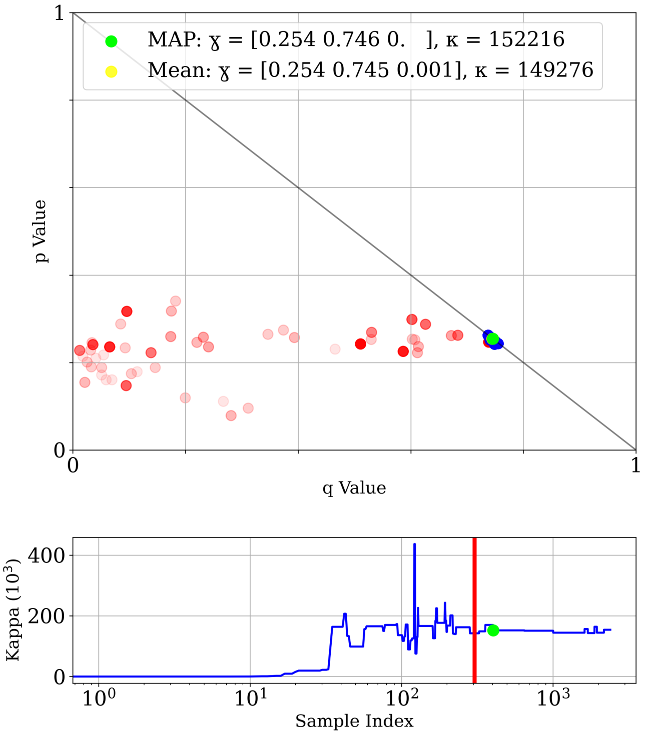


Figure 4.14: Metropolis Hastings samples for mixing ratio and Kappa value

For our approach, we will use the three introduced single hypotheses and determine an optimal mixing ratio $\gamma = [p, q, 1 - p - q]$, where p, q and $1 - p - q$ denote the percentages to explain transitions according to the hub hypothesis, similarity hypothesis or the link hypothesis, respectively in that order. As far as the hypothesis conception goes, our approach stands in contrast to MixedTrails. While their (mixed) hypothesis deterministically assigns different (single) hypotheses to certain parts of a trail, our approach yields group assignment probabilities, that are used for every transition of a trail, i.e. for the whole data set.

Figure 4.14 shows the results for a Metropolis Hastings run with 2.400 iterations. We used a flat Dirichlet prior over the space of mixing ratios and an Exponential distribution with expected value of 100.000 over the space of Kappa values. As before, this expected value is chosen in accordance with the number of transitions the data set consists of. The proposal distributions are a Dirichlet distribution with concentration factor $c = 100$ to up scale the current mixing ratio, which is taken as a parameter and an Exponential distribution, which takes the current Kappa value as an expectancy. We initialized Kappa to have a value of 10. In this run 61 samples were accepted. It takes about 300 iterations to obtain samples, that cluster around a mixing ratio of $\gamma = [0.25, 0.75, 0]$, while simultaneously reaching a Kappa value of ~ 150.000. The fact, that the Kappa value seems to settle around this value, coincides with our choice of prior and the size of the data set. As for the mixing ratio, the MAP and mean both yield approximately $\gamma = [0.25, 0.75, 0]$ as a parameter. To obtain a maximal evidence, these estimates suggest to explain transitions with a 25% probability according to the hub hypothesis, with 75% according to the similarity hypothesis and to explain none of them according to the link hypothesis. The last interpretation, that this approach proposes to never explain transitions according to the link hypothesis, is noticeably. This means, that the two hypotheses - hub and similarity - yield a better explainability and that users actually navigate according to a mixture of both of them. This and the received ratio of $[0.25, 0.75]$ for the hub and similarity hypotheses are in line with the results presented by MixedTrails [2]. They obtained the highest evidence, when they assigned the hub hypothesis to first two transitions of a trail and the similarity hypothesis to the rest of each trail. Given an average trail length of 5.6, this would - even though both approaches are not directly comparable as they act on different levels - give probabilities of roundabout 0.36 to choose the hub and 0.64 to choose the similarity hypothesis. However, we can compare the evidence values, since the data set is identical. For our approach, as presented in figure 4.14, we obtain a maximal evidence of -511.105 for a Kappa value of 152.216 and the mentioned mixing ratio of $\gamma = [0.25, 0.75, 0]$. The approach taken by MixedTrails yields a maximal evidence of -504.631 for a Kappa value of 100.000 using the mixed hypothesis combining the hub (for the first two clicks) and the similarity (for the remaining clicks) hypothesis. For this data set, the hypothesis conception carried out in the MixedTrails paper, yields a slightly higher evidence and as a result, gives a better explanation of the observed trails. Nevertheless, the tendencies of the two approaches are similar. Firstly, that both - the hub and similarity hypothesis - should be preferred over the link hypothesis and secondly, how the ratio of hub and similarity hypothesis should approximately be chosen.

5 Conclusion

In this thesis, we proposed an approach to find optimal compositions of hypotheses for sequential data using Bayesian inference. As a result, our method contributes to answering the question of finding mixing ratios of hypotheses that explain the data best. We coupled existing theory and frameworks for hypothesis comparison to the field of inference via Markov Chain Monte Carlo methods, opening up the existing approaches to a different application area.

5.1 Summary and Contributions

This work consists of three main parts. In the first part, we have introduced HypTrails and MixedTrails, the two main building blocks that this work connects to. We showed how these two approaches formulate hypotheses and compare them using first order Markov chain modelling and Bayes factors. Of particular importance for us was MixedTrails and their approach to formulate compositions or mixtures of hypotheses. Our method, which was presented in the next chapter, used the Metropolis Hastings algorithm, a Markov Chain Monte Carlo algorithm, to generate approximate posterior samples. These samples were then used to simulate the corresponding distribution of mixing ratios and obtain optimal compositions with respect to the maximum a posteriori (MAP) or mean estimator. We established the necessary theory for Markov chains on general state spaces, which is needed to proof convergence of the algorithm and showed that all properties are indeed fulfilled in all our experimental settings. The last chapter was devoted to carrying out experiments for several scenarios. We employed the Metropolis Hastings algorithm to a variety of synthetic data sets with distinctive hypothesis conceptions, as well as for the empirical Wikispeedia data set and demonstrated its applicability.

5.2 Methodology and Outlook

With our method, we were able to find optimal hypotheses compositions by imposing a prior over the space of mixing ratios. To be able to this, throughout the inference process, we use the same group assignment probabilities for every transition in the data set. This means, that we take away the possibility to identify behavioural changes within a trail. Of course, opening up the parameter space to distinguishing between every single observed transition is computationally not feasible. However, being able to identify such trails, where a change of transitional behaviour can be observed, is desirable, as we have seen in the last section, where, for the Wikispeedia dataset, a split of hypotheses held a higher evidence than a mixture.

The experiments, that we have carried out, did not reach the limit of the capability of our method in terms of computational cost and execution time of the algorithm. There is some room to expand the parameter space to higher dimensions and still be able to extract maximum evidence estimates. How much this is precisely, requires additional analysis. When further expanding the parameter space, at some point, employing more complex MCMC algorithms, such as the Gibbs sampler, which produces a sequence of low dimensional approximations, would become necessary. On our behalf, to demonstrate the substantial functionality of our approach while still being able to clearly interpret the result, was the desired result for this work.

Bibliography

[1] Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15. International World Wide Web Conferences Steering Committee, May 2015.

[2] Martin Becker, Florian Lemmerich, Philipp Singer, Markus Strohmaier, and Andreas Hotho. Mixedtrails: Bayesian hypotheses comparison on heterogeneous sequential data. *Data Mining and Knowledge Discovery*, 31, 09 2017.

[3] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[4] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.

[5] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1(none), January 2004.

[6] S. Rosenthal. A review of asymptotic convergence for general state space markov chains. *Far East Journal of Theoretical Statistics*, 5, 01 2001.

[7] Patrick Billingsley. *Probability and measure.* A Wiley-Interscience publication. Wiley, New York [u.a.], 3. ed edition, 1995.

[8] Robert West and Jure Leskovec. Human wayfinding in information networks. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 619–628, New York, NY, USA, 2012. Association for Computing Machinery.

[9] Martin Becker. *Understanding Human Navigation using Bayesian Hypothesis Comparison.* Doctoral thesis, University of Würzburg, 2018.

[10] J.L. Doob. *Stochastic Processes*. Probability and Statistics Series. Wiley, 1953.

[11] K. B. Athreya and P. Ney. A new approach to the limit theory of recurrent markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978.

[12] Luke Tierney. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701 – 1728, 1994.

[13] Gareth O. Roberts and Jeffrey S. Rosenthal. Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. *The Annals of Applied Probability*, 16(4), November 2006.

[14] Krishna B. Athreya and Peter Ney. A new approach to the limit theory of recurrent markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978.

[15] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[16] Martin Becker, Philipp Singer, Florian Lemmerich, Andreas Hotho, Denis Helic, and Markus Strohmaier. Photowalking the city: Comparing hypotheses about urban photo trails on flickr. pages 227–244, 12 2015.

[17] Martin Becker, Hauke Mewes, Andreas Hotho, Dimitar Dimitrov, Florian Lemmerich, and Markus Strohmaier. Sparktrails: A mapreduce implementation of hyptrails for comparing hypotheses about human trails. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 17–18, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.