Continuous Hidden Markov Models and the Sequential Probability Ratio Test

Oscar Darwin

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

Acknowledgements

I would like to thank my supervisor Stefan Kiefer. He remained logically steadfast throughout my many incorrect ideas and wild tangents. He put me on the right track when I was exploring in the wrong direction. Finally, he helped me grow important ideas and translate them into concrete language. It has been a delight working together.

I would like to acknowledge the financial support of the Royal Society who helped to fund this research.

Thank you to Bernard Sufrin for being such a fervent supporter of mine from the moment I came to Oxford. Further, I would like to thank the members and associates of the Bartlemas Institute for the many inspirational sessions, out of which so many of the ideas in this thesis sprung.

Furthermore, I would like to thank my family: Lucy, Andrew and Frances for giving me a place to rest my mind during particularly stressful and difficult problems. In the final moments of my research and writing up, Lila encouraged me with impeccable patience and warmth. Thank you for being there at the finish line.

Abstract

We consider Hidden Markov Models that emit sequences of observations which are drawn from continuous distributions. For example, such a model may emit a sequence of numbers, each of which is drawn from a uniform distribution, but the support of the uniform distribution depends on the state of the Hidden Markov Model. Such models generalise the more common version where each observation is drawn from a finite alphabet. We consider a distance measure on Hidden Markov Models called the total variation distance. When this distance is 0 we say the models are equivalent. When this distance is maximal we say the models are distinguishable. We prove that for two Hidden Markov Models with continuous observations one can decide in polynomial time whether they are equivalent and also whether they are distinguishable. We also consider the Sequential Probability Ratio Test applied to Hidden Markov Models with finite observations. Given two distinguishable Hidden Markov Models and a sequence of observations generated by one of them, the Sequential Probability Ratio Test attempts to decide which model produced the sequence. We show relationships between the execution time of such an algorithm and Lyapunov exponents of random matrix products. Further, we give complexity results about the execution time taken by the Sequential Probability Ratio Test.

Contents

1
Introduction

Two computer programs run simultaneously. We see terminal outputs of the programs and obtain a glimpse into the underlying logic. Is it possible to tell the programs apart? Can we always tell them apart? If we are shown just one of the terminal outputs, how quickly can we decide which program the terminal output belongs to?

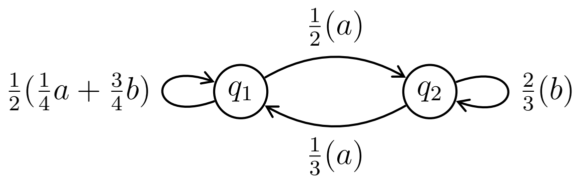
In the real world, randomness in the inputs to programs, randomness in the computer circuitry and in the case of concurrent processes, randomness in the ordering of operations may lead to probabilistic execution of the programs. By carefully modelling the input randomness, we can model the execution randomness too. So, the questions in the first paragraph can be generalised to questions about the probability distributions of the visible trace of the two programs.

A *Hidden Markov Model* (HMM) aims to model such a program. The model is a state-based transition model and so captures the discreteness of a program as it evolves over time. A program typically has many states and we see only partial information on its execution. A Hidden Markov model is similar in that it emits a signal each time the state changes.

The trace of a program need not be discrete. Consider a program where every command takes a different amount of time to execute. Moreover, suppose we know the start and end time of each command. Then, the visible trace of the program can be modelled as a positive real number. With the execution time of each command we can infer which command is being executed. Knowledge about the internal workings of a program is a potential security vulnerability. Is it possible to tell two programs apart based on just the timing of particular commands? Can we always tell them apart? How quickly can an attacker infer which of two such programs are running?

1.1 Hidden Markov Models

A (discrete-time, finite-state) Hidden Markov Model (often called *labelled Markov chain*) has a finite set Q of states and for each state a probability distribution over its possible successor states. For any two states q, q', whenever the state changes from q to q', the HMM samples and then emits a random observation according to a probability distribution $D(q, q')$. For example, consider the following diagram visualising a HMM:



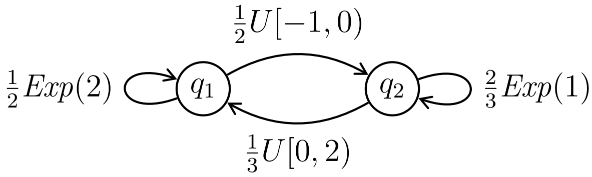
In state q_1, the successor state is q_1 or q_2, with probability $\frac{1}{2}$ each. Upon transitioning from q_1 to itself, observation a is drawn with probability $\frac{1}{4}$ and observation b is drawn with probability $\frac{3}{4}$; upon transitioning from q_1 to q_2, observation a is drawn surely.[1]

In this way, a HMM, together with an initial distribution on states, generates a random infinite sequence of observations. In the example above, if the initial distribution is the Dirac distribution on q_1, the probability that the observation sequence starts with a is $\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2}$ and the probability that the sequence starts with ab is $\frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{2}{3}$.

In the example above the observations are drawn from a finite set of observations $\Sigma = \{a, b\}$. Indeed, in the literature HMMs most commonly have a finite set of observations. In this thesis we lift this restriction and consider *continuous-observation* HMMs, by which we mean HMMs as described above, but with an infinite observation set Σ. For example, instead of the distributions on $\{a, b\}$ in the picture above (written there as $(\frac{1}{4}a + \frac{3}{4}b)$, (a), (b), respectively), we may have distributions on the real numbers. An example is the following diagram,

[1] One may allow for observations also on the states and not only on the transitions. But such state observations can be equivalently emitted upon leaving the state. Hence we can assume without loss of generality that all observations are emitted on the transitions.

where $U[a,b)$ denotes the uniform distribution on $[a,b)$ and $Exp(\lambda)$ denotes the exponential distribution with parameter λ.

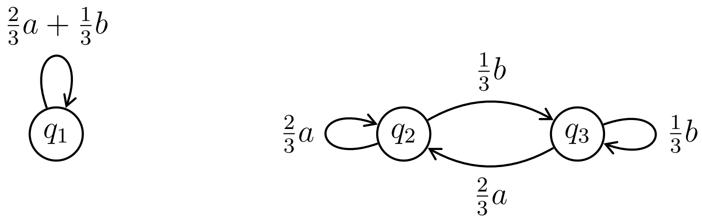


Continuous-observation HMMs also generate a random infinite sequence of observations. For example, starting with initial state q_1, the model transitions to q_1 and q_2 both with probability $\frac{1}{2}$. If it takes the transition to q_1, the observation is drawn from an exponential distribution with rate 2. If it takes the transition to q_2, the observation is drawn from a uniform distribution on the half-open interval $[-1,0)$. This process continues to generate the random infinite sequence. For example, such a sequence could be $1.2, \pi, -\frac{1}{2}, 2, 1, \ldots$ and so on.

1.2 Equivalence and Distinguishability

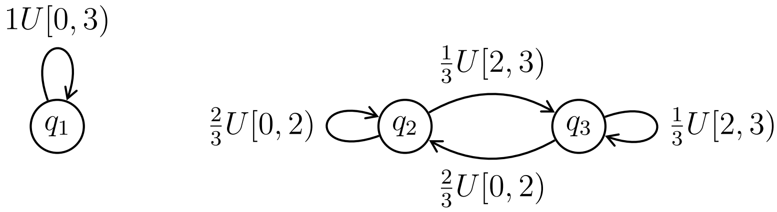
HMMs, both with finite and infinite observation sets, are widely employed in fields such as speech recognition (see [1] for a tutorial), gesture recognition [2], signal processing [3], and climate modeling [4]. HMMs are heavily used in computational biology [5], more specifically in DNA modeling [6] and biological sequence analysis [7], including protein structure prediction [8] and gene finding [9]. In computer-aided verification, HMMs are the most fundamental model for probabilistic systems; model-checking tools such as Prism [10] and Storm [11] are based on analyzing HMMs efficiently.

One of the most fundamental questions about HMMs is whether two HMMs with initial state distributions are *(trace) equivalent*, i.e., generate the same distribution on infinite observation sequences. Consider the two HMMs below. At states q_2 and q_3 the probability of producing a or b is the same as in state q_1. Thus the probability of any given sequence of observations is the same whether one starts the HMM from q_1 or q_2. In the example below, (the Dirac distributions on) states q_1 and q_2 are equivalent since each state has the same distribution on observations.



For finite observation alphabets this problem is very well studied and can be solved in polynomial time using algorithms that are based on linear algebra [12–15]. Checking trace equivalence is used in the verification of obliviousness and anonymity, properties that are hard to formalize in temporal logics, see, e.g., [16–18].

We consider trace equivalence in the context of continuous-observation HMMs. Unlike with finite-observation HMMs, it is possible for a continuous pdf to be simulated by the combination of multiple other continuous pdfs. Consider the HMMs below where $U[a, b)$ denotes the uniform distribution on the interval $[a, b)$.



From both states q_2 and q_3 the combination of choosing a uniform random number in $[0, 2)$ with probability $\frac{2}{3}$ and a uniform random number in $[2, 3)$ with probability $\frac{1}{3}$ is the same as choosing a uniform random number in $[0, 3)$. Thus the Dirac distributions on the states q_1 and q_2 are equivalent.

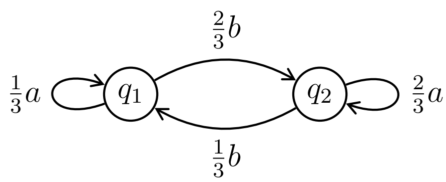
Although the generalisation to continuous observations (such as passed time, consumed energy, sensor readings) is natural, there has been little work on the algorithmics of such HMMs. One exception is *continuous-time* Markov chains (CTMCs) [19, 20] which are similar to HMMs described above, but with two kinds of observations: on the one hand they emit observations representing the current state of the CTMC, but on the other hand they also emit the *time* spent in each state. Typically, each state-to-state transition is labelled with a parameter λ; for each transition its time of "firing" is drawn from an exponential distribution with parameter λ; the transition with the smallest firing time "wins" and causes the corresponding change of state. CTMCs have attractive properties: they are in a sense memoryless, and for many analyses, including model checking, an equivalent discrete-time model can be calculated using an efficient and numerically stable process called *uniformization* [21].

In [22] a stochastic model more general than ours was introduced, allowing not only for uncountable sets of observations (called *labels* there), but also for infinite sets of states and actions. The paper [22] focuses on bisimulation; trace equivalence is not considered. It emphasizes nondeterminism, a feature we do not consider here.

Equivalence is a strong notion, and a natural question about nonequivalent distributions in a given HMM is *how* different they are. For initial distributions π_1, π_2 on the states of the HMM, let us write $\mathbb{P}_{\pi_1}, \mathbb{P}_{\pi_2}$ for the induced probability measure on infinite observation sequences; i.e., $\mathbb{P}_{\pi_i}(E)$, for a measurable event $E \subseteq \Sigma^\omega$, is the probability that the random infinite word $w \in \Sigma^\omega$ produced starting from π_i is in E. Then, the *total variation distance* between $\mathbb{P}_{\pi_1}, \mathbb{P}_{\pi_2}$ is defined as

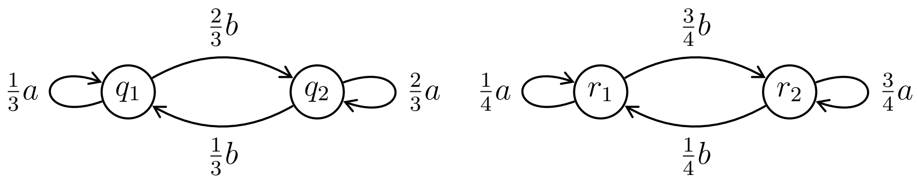
$$d(\pi_1, \pi_2) \ := \ \sup \left\{ |\mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)| \mid \text{measurable } E \subseteq \Sigma^\omega \right\}.$$

This supremum is a maximum; i.e., there always exists a "maximizing event" $E \subseteq \Sigma^\omega$ with $d(\pi_1, \pi_2) = \mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)$. In these terms, initial distributions π_1, π_2 are equivalent if and only if $d(\pi_1, \pi_2) = 0$. The total variation distance was studied in more detail in [23]. There it was shown that for finite observation HMMs, the problem whether $d(\pi_1, \pi_2) = 1$ holds can also be decided in polynomial time. Call distributions π_1, π_2 *distinguishable* if $d(\pi_1, \pi_2) = 1$. Distinguishability was used for runtime monitoring [24] and diagnosability [25, 26] of stochastic systems. In the example below, (the Dirac distributions on) states q_1 and q_2 are distinguishable.



Two initial distributions π_1, π_2 that are distinguishable (i.e., $d(\pi_1, \pi_2) = 1$) can nevertheless be "hard" to distinguish. If we replace the transition probabilities $\frac{1}{3}, \frac{2}{3}$ in the HMM by $\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon$, respectively, states q_1, q_2 remain distinguishable for every $\varepsilon > 0$, although, intuitively, the smaller $\varepsilon > 0$ the more observations are needed to define an event E such that $\mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)$ is close to 1.

As a concrete example, consider below the HMMs for $\varepsilon = \frac{1}{6}$ and $\varepsilon = \frac{1}{4}$ on the left and right respectively. It takes more observations to distinguish q_1 and q_2 than it does to distinguish r_1 and r_2.



1.3 Likelihood Ratios

For initial distributions π_1, π_2, a word $w \in \Sigma^\omega$ and $n \in \mathbb{N}$ consider the *likelihood ratio*

$$L_n(w) \;:=\; \frac{\mathbb{P}_{\pi_1}(w_n\Sigma^\omega)}{\mathbb{P}_{\pi_2}(w_n\Sigma^\omega)}\,,$$

where w_n denotes the length-n prefix of w. In the left HMM in the example above, we have $\mathbb{P}_{q_1}(aba\Sigma^\omega) = \frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}$ and $\mathbb{P}_{q_2}(aba\Sigma^\omega) = \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{3}$. Thus, for any word w starting with aba we have $L_n(w) = 2$. We consider the likelihood ratio L_n as a random variable for every $n \in \mathbb{N}$. It turns out more natural to focus on the *log-likelihood ratio* $\ln L_n$. One can show that the limit $\lim_{n\to\infty} \ln L_n \in [-\infty, \infty]$ exists $\mathbb{P}_{\pi_1}$-almost surely and $\mathbb{P}_{\pi_2}$-almost surely (see, e.g., [23, Proposition 6]). In fact, if π_1, π_2 are distinguishable, then $\lim_{n\to\infty} \ln L_n = \infty$ holds $\mathbb{P}_{\pi_1}$-almost surely and $\lim_{n\to\infty} \ln L_n = -\infty$ holds $\mathbb{P}_{\pi_2}$-almost surely. This suggests the "average slope", $\lim_{n\to\infty} \frac{1}{n} \ln L_n$, of $\ln L_n$ as a measure of *how* distinguishable two distinguishable distributions π_1, π_2 are.

In Figure 1.1, we compare the log-likelihood ratios of both pairs of initial distributions in the previous pair of HMMs. The log-likelihood started from the Dirac distributions on r_1 and r_2 has a steeper slope than the log-likelihood started from the Dirac distributions on q_1 and q_2.

The log-likelihood ratio plays a central role in the *sequential probability ratio test (SPRT)* [27], which is optimal [28] among sequential hypothesis tests (such tests attempt to decide between two hypotheses without fixing the sample size in advance). In terms of a HMM and two initial distributions π_1, π_2, the SPRT attempts to

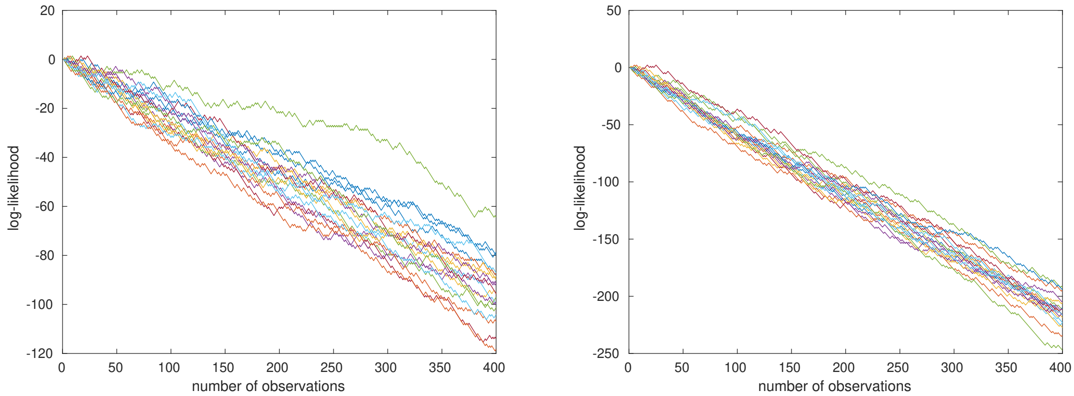


Figure 1.1: The two images show two log-likelihood plots. The left plot is the log-likelihood started from the Dirac distributions on q_1 and q_2 with observations produced by q_2. The right plot shows the same, but for r_1 and r_2. Both plots show 20 sample runs with 400 observations.

decide, given longer and longer prefixes of an observation sequence $w \in \Sigma^\omega$, which of π_1, π_2 is more likely to emit w. The SPRT works as follows: fix a lower and an upper threshold (which determine type-I and type-II errors); given increasing prefixes of w keep track of $\ln L_n(w)$, and when the upper threshold is crossed output π_1 and stop, and when the lower threshold is crossed output π_2 and stop. Again, it is natural to assume that the average slope of increase or decrease of $\ln L_n$ determines how long the SPRT needs to cross one of the thresholds.

For example, suppose we choose lower and upper thresholds -40 and 40 respectively. We see that all sample runs in both plots in Figure 1.1 cross the lower threshold before the upper threshold. Therefore, the SPRT terminates for all sample runs and ouputs q_2 and s_2 respectively for each plot. In the left plot, the SPRT terminates after approximately 110 observations in the best case and after approximately 330 observations in the worst case. In the right plot, the best case is approximately 50 observations and the the worst case is approximately 100 observations. Thus, on average it takes at least twice as long for the SPRT to terminate for the left plot. Indeed, the slope of the right plot is more than twice as steep as the slope of the left plot.

If the average slope $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ exists and equals a number ℓ with positive probability, we call ℓ a *likelihood exponent*. The term is motivated by a close relationship to *Lyapunov exponents*, which characterise the growth rate of certain random matrix products. In Figure 1.2 we show the average slopes of the plots in

Figure 1.1. In this case, there is exactly one likelihood exponent for each. As we explain in Section 7.1, since both HMMs are deterministic, the likelihood exponents can be computed as $-\frac{1}{3}\ln 2$ and $-\frac{1}{2}\ln 3$ respectively. In general there can be more than one likelihood exponent: see Section 7.2. Moreover, these likelihood exponents can be hard to compute, although there are approximation algorithms using convex optimisation [29, 30].

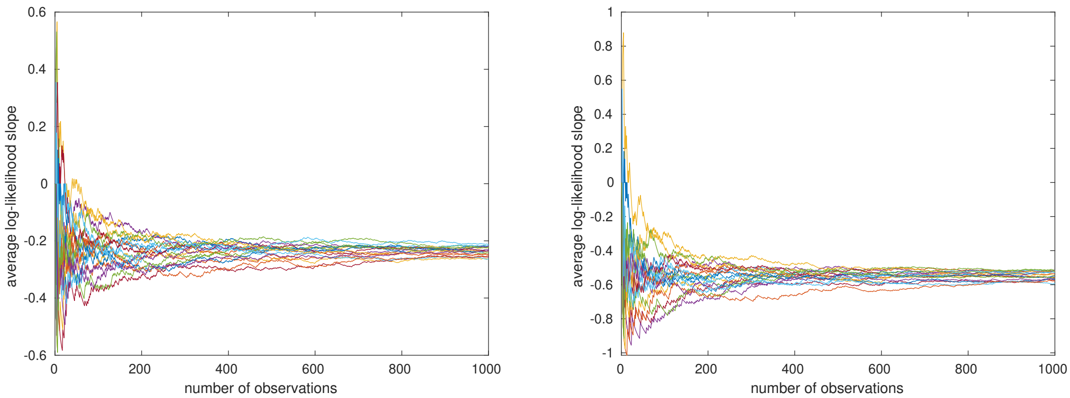


Figure 1.2: The two images show the average slopes of the respective log-likelihood plots in Figure 1.1 for a given number of observations. Both plots show 20 sample runs with 1000 observations.

1.4 Symbolic Computation

In the case of finite-observation HMMs, it is straightforward to specify the probability distributions on each transition. When the set of observations is infinite, symbolic representations of the pdfs are required because we need to reason about the pdfs computationally. For example we saw earlier that a continuous pdf can be simulated by multiple other continuous pdfs. This happens if there is a linear dependence between the pdfs. Our equivalence algorithm must be able to determine any linear dependence between the pdfs using the symbolic representations. Similarly our distinguishability algorithm must be able to determine whether the supports of any two pdfs overlap. We encode the probability density functions on their transitions using *profiles*. Each profile is a finite string consisting of the parameters of the pdf which we assume to be rational. For example, the uniform distribution $U[a, b)$

is encoded with its end points $a, b \in \mathbb{Q}$. The exponential distribution $Exp(\lambda)$ is encoded with just its rate parameter λ.

1.5 Research Questions

This thesis focusses on the comparison of two HMMs and is grouped into *comparison questions*.

- **Q1:** How can we encode continuous-observation HMMs in such a way that one can efficiently check trace equivalence?
- **Q2:** How can we encode continuous-observation HMMs in such a way that one can efficiently check whether two HMMs are distinguishable?
- **Q3:** How many observations are required to distinguish two HMMs?

1.6 Structure of this Document

We have ordered this document broadly by the comparison questions above. It follows a natural sequence. We begin by looking at the similarity of HMMs through the lens of probabilistic equivalence. After that, we introduce the total variation distance as a way to measure the difference of two HMMs. We then show that even if two HMMs are maximally probabilistically different, i.e. distinguishable, they can take different numbers of observations to distinguish via the SPRT.

- In Chapter 3 we show an equivalence algorithm for continuous-observation HMMs. Towards the first research question, we introduce *linearly decomposable profile languages* to encode continuous-observation HMMs and show that if a HMM is encoded with a linearly decomposable profile language one can decide whether two initial distributions are equivalent in polynomial time.
- In Chapter 4 we show that the algorithm for deciding whether two initial distributions are distinguishable in finite-observation HMMs has a continuous-observation analogue. Towards the second research question, we introduce joint-graph decomposable profile languages and show that if a HMM is encoded with a joint-graph decomposable and linearly decomposable profile language

then one can decide in polynomial time whether two initial distributions are distinguishable.

- In Chapter 5 we define a profile language with which a large class of standard probability density functions can be encoded. We show this profile language is linearly decomposable and joint-graph decomposable. We also show that the set of observations does not need to be one-dimensional. We define a profile language for encoding Dirichlet distributions and show that this profile language is also linearly decomposable and joint-graph decomposable.

- In Chapter 6 we show that even if two initial distributions are distinguishable, their likelihood ratio may exhibit different convergence rates. In particular, for finite-observation HMMs we exhibit a tight connection between the SPRT and likelihood exponents; i.e., the time taken by the SPRT depends on the likelihood exponents of the HMM. The content in this chapter addresses the third research question.

- In Chapter 7 we show that for finite-observation *deterministic* HMMs one can compute likelihood exponents in polynomial time. Further, for finite-observation HMMs we show that in general the likelihood exponents can be efficiently expressed in terms of Lyapunov exponents. Moreover, We show that the likelihood exponent exists almost surely and prove a quadratic bound on the number of likelihood exponents.

1.7 Contributions

The majority of Chapter 3 and some elements of Chapter 5 relating to linearly decomposable profile languages were published in FSTTCS 2020. See [31]. The contents of Chapter 6 and Chapter 7 where published in CONCUR 2022. See [32]. The remaining material in Chapter 4 and Chapter 5 builds on Chapter 3.

2 Definitions

2.1 Notations

We write $\mathbb{N}$ for the set of positive integers, $\mathbb{N}_0$ for the set of non-negative integers, $\mathbb{Q}$ for the set of rationals and $\mathbb{Q}_+$ for the set of positive rationals. For $d \in \mathbb{N}$ and a finite set Q we use the notation $|Q|$ for the number of elements in Q, $[d] = \{1, \ldots, d\}$ and $[Q] = \{1, \ldots, |Q|\}$. Vectors $\mu \in \mathbb{R}^N$ are viewed as row vectors. The norm $\|\mu\|$ is assumed to be the l_1 norm: $\|\mu\| = \sum_{q \in Q} |\mu_q|$. We often identify vectors $\mu \in [0,1]^Q$ such that $\|\mu\| = 1$ with the corresponding probability distribution on Q.

We write $\overline{0}, \overline{1}$ for the vectors all of whose entries are 0, 1, respectively. Superscript $\top$ denotes transpose; e.g., $\overline{1}^\top$ is a column vector of ones. For $q \in Q$, we denote by $e_q \in \{0,1\}^Q$ the vector with $(e_q)_q = 1$ and $(e_q)_{q'} = 0$ for $q' \neq q$.

A matrix $M \in \mathbb{R}^{N \times N}$ is *stochastic* if M is non-negative and $\sum_{j=1}^{N} M_{i,j} = 1$ for all $i \in [N]$. For a domain Σ and subset $E \subseteq \Sigma$ the *characteristic* function $\mathbb{1}_E : \Sigma \to \{0,1\}$ is defined as $\mathbb{1}_E(x) = 1$ if $x \in E$ and $\mathbb{1}_E(x) = 0$ otherwise.

Throughout this thesis, we use Σ to denote a set of *observations*. We assume Σ is a topological space and $(\Sigma, \mathscr{G}, \lambda)$ is a measure space where all the open subsets of Σ are contained within $\mathscr{G}$ and have non-zero measure. Indeed $\mathbb{R}$ and the usual Lebesgue measure space on $\mathbb{R}$ satisfy these assumptions. For a finite alphabet Σ and $n \in \mathbb{N}$ we denote by $\Sigma^n, \Sigma^*, \Sigma^+, \Sigma^\omega$ the sets of length-n words, finite words, non-empty finite words, infinite words, respectively. For $w \in \Sigma^\omega$ we write w_n for the length-n prefix of w.

A matrix valued function $\Psi : \Sigma \to [0,\infty)^{N \times N}$ can be integrated element-wise. We write $\int_E \Psi \, d\lambda$ for the matrix with entries $(\int_E \Psi \, d\lambda)_{i,j} = \int_E \Psi_{i,j} \, d\lambda$, where $\Psi_{i,j} : \Sigma \to [0,\infty)$ is defined by $\Psi_{i,j}(x) = \big(\Psi(x)\big)_{i,j}$ for all $x \in \Sigma$. For a row vector $\mu \in [0,\infty)^Q$ we write $\operatorname{supp}(\mu) := \{q \in Q \mid \mu_q > 0\}$. For a matrix $A \in [0,\infty)^{Q \times Q}$ we write $\operatorname{supp}(A) = \{(q,r) \in Q \times Q \mid A_{q,r} > 0\}$. For a function $f : \Sigma \to \mathbb{R}$ we write $\operatorname{supp}(f) = \{x \in \Sigma \mid f(x) \neq 0\}$.

Consider a function $f : \Sigma \to \mathbb{R}^m$. Let $\mathring{\Sigma} \subseteq \Sigma$ be the set of all points at which f is continuous. We refer to Σ as the *set of continuity*. If $\lambda(\mathring{\Sigma}) = \lambda(\Sigma)$ we say that f is *almost-surely continuous*.

Definition 2.1.1. A *Hidden Markov Model* (HMM) is a triple (Q, Σ, Ψ) where Q is a finite set of states, Σ is a set of observations, and the *observation density matrix* $\Psi : \Sigma \to [0,\infty)^{|Q| \times |Q|}$ specifies the transitions such that $\int_\Sigma \Psi \, d\lambda$ is a stochastic matrix. We say a HMM is a continuous-observation HMM if Ψ is almost-surely continuous.

Example 2.1.1. The second HMM from the introduction is the triple $(\{q_1, q_2\}, \mathbb{R}, \Psi)$ with

$$\Psi(x) \;=\; \begin{pmatrix} \frac{1}{2} \cdot 2\exp(-2x) \cdot \chi_{[0,\infty)}(x) & \frac{1}{2} \cdot 1 \cdot \chi_{[-1,0)}(x) \\ \frac{1}{3} \cdot \frac{1}{2} \cdot \chi_{[0,2)}(x) & \frac{2}{3} \cdot \exp(-x) \cdot \chi_{[0,\infty)}(x) \end{pmatrix}. \qquad \square$$

Fix a HMM $\mathcal{H} = (Q, \Sigma, \Psi)$ for the rest of the section. We extend Ψ to the mapping $\Psi : \Sigma^* \to [0,1]^{Q \times Q}$ with $\Psi(a_1 \cdots a_n) = \Psi(a_1) \cdot \ldots \cdot \Psi(a_n)$ and $\Psi(\varepsilon) = I$, where ε is the empty word and I the $Q \times Q$ identity matrix.

2.2 Paths, Runs and Cylinder Sets

We call a finite sequence $v = q_0 a_1 q_1 \cdots a_n q_n \in Q(\Sigma Q)^*$ a *path* and an infinite sequence $q_0 a_1 q_1 a_2 q_2 \cdots \in Q(\Sigma Q)^\omega$ a *run*. We say that $A \subseteq Q(\Sigma Q)^n$ is a *cylinder set* if $A = \{q_0\} \times A_1 \times \{q_1\} \times \cdots \times A_n \times \{q_n\}$ where $A_i \in \mathscr{G}$ for $i = 1, \ldots, n$ and $q_i \in Q$ for $i = 0, \ldots, n$. We sometimes abbreviate cylinder sets as $q_0 A_1 q_1 \cdots A_n q_n$.

To $\mathcal{H}$ and an *initial probability distribution* $\pi \in [0,1]^Q$ we associate the probability space $(Q(\Sigma Q)^\omega, \mathscr{G}^\omega\!\uparrow, \mathbb{P}_\pi)$ where $\mathscr{G}^\omega\!\uparrow$ is the σ-algebra generated by the cylinder sets and $\mathbb{P}_\pi$ is the unique probability measure with $\mathbb{P}_\pi(q_0 A_1 q_1 \cdots A_n q_n (\Sigma Q)^\omega) = \pi_{q_0} \int_{A_i} \Psi_{q_0,q_1} d\lambda \times \cdots \times \int_{A_n} \Psi_{q_{n-1},q_n} d\lambda$.

We refer to $A_1 \times \cdots \times A_n$ where $A_i \in \mathscr{G}$ for all $i \in [n]$ as an *observation* cylinder set. We define the measure space $(\Sigma^n, \mathscr{G}^n, \lambda^n)$ where $\mathscr{G}^n$ is the smallest σ-algebra containing the observation cylinder sets of length n. If $\mathring{\Sigma}$ is the set of continuity for $\Psi : \Sigma \to [0, \infty)^{|Q| \times |Q|}$, then the extension $\Psi : \Sigma^n \to [0, \infty)^{|Q| \times |Q|}$ is almost-surely continuous with respect to λ^n with set of continuity $\mathring{\Sigma}^n$.

We define $\mathscr{G}^\omega$ as the smallest σ-algebra containing the observation cylinder sets. As the states are often irrelevant, for $E \in \mathscr{G}^\omega$ and $E\uparrow := \{q_0 a_1 q_1 a_2 q_2 \cdots \mid a_1 a_2 \cdots \in E\} \in \mathscr{G}^\omega\uparrow$ we also view E as an event and may write $\mathbb{P}_\pi(E)$ to mean $\mathbb{P}_\pi(E\uparrow)$. In particular, for an observation cylinder set $A_1 \times \cdots \times A_n$ we have $\mathbb{P}_\pi(A_1 \times \cdots \times A_n \Sigma^\omega) = \|\pi \int_{A_1} \Psi d\lambda \times \int_{A_n} \Psi d\lambda\|$. By $\mathbb{E}_\pi$ we denote the expectation with respect to $\mathbb{P}_\pi$. If π is the Dirac distribution on state q, then we write $\mathbb{E}_q$.

2.3 Markov Chains

A *Markov chain* is a pair (Q, T) where Q is a finite set of states and $T \in [0, 1]^{Q \times Q}$ is a stochastic matrix. A Markov chain (Q, T) is naturally associated with its directed *graph* $(Q, \{(q, r) \mid T_{q,r} > 0\})$, and so we may use graph concepts, such as strongly connected components (SCCs), in the context of a Markov chain. Trivial SCCs are considered SCCs. For a directed graph $G = (V, E)$ and $U, U' \subseteq V$ we say $U \to_G U'$ if there is $x \in U$ and $y \in U'$ such that y is reachable from x in G.

A Markov chain (Q, T) and an initial distribution $\iota \in [0, 1]^Q$ are associated with a probability measure $\mathbb{P}_\iota$ on measurable subsets of Q^ω; the construction of the probability space is similar to HMMs, without the observation alphabet Σ.

The *embedded* Markov chain of a HMM (Q, Σ, Ψ) is the Markov chain $(Q, \int_\Sigma \Psi d\lambda)$. We say that a HMM is strongly connected if the graph of its embedded Markov chain is strongly connected.

Example 2.3.1. The embedded Markov chain of the HMM from Example 2.1.1 is $(\{q_1, q_2\}, \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix})$.

The embedded graph of a HMM is the embedded graph of its embedded Markov chain.

2.4 Finite-Observation HMMs

Many of the definitions stated for continuous-observation HMMs have finite analogues. Let Σ be a finite set of observations and let λ be the counting measure. We say a HMM (Q, Σ, M) is a finite-observation HMM if Σ is a finite set and $\int_\Sigma M d\lambda = \sum_{a\in\Sigma} M(a)$ is stochastic. The *embedded* Markov chain of a finite-observation HMM (Q, Σ, M) is the Markov chain $(Q, \sum_{a\in\Sigma} M(a))$. When Σ is finite, we assume for computational purposes that the numbers in M are rational and expressed as fractions of integers encoded in binary. Computation complexity is assessed in the Turing model and so testing for equality of members takes linear time. In the case of a finite-observation HMM we may refer to the cylinder set $\{q_0\} \times \{a_1\} \times \{q_1\} \times \cdots \times \{a_n\} \times \{q_n\}$ as simply $q_0 a_1 q_1 \cdots a_n q_n$.

3
Equivalence

In this chapter we explore equivalence in continuous-observation HMMs. We show in Section 3.1 that certain aspects of the linear-algebra based approach for checking equivalence of finite-observation HMMs carry over to the continuous case naturally. In particular, equivalence reduces to orthogonality in a certain vector space of state-indexed real vectors: see proposition 3.1.1.

However, we show in section 3.2 that in the continuous case there can be *additional* linear dependencies between the observation density functions (which is impossible in the finite case, where the different observations can be assumed linearly independent). This renders a simple-minded reduction to the finite case incorrect. Therefore, an equivalence checking algorithm needs to consider the vector space from the previous paragraph.

For the required computations on the observation density functions we introduce in section 3.3 *linearly decomposable profile languages*, which are languages (i.e., sets of finite words) whose elements encode density functions on which basis computations can be performed efficiently. In section 3.3 we provide an example of such a language, encoding (linear combinations of) Gaussian, exponential, and piecewise polynomial density functions.

In section 3.5 we finally show that HMMs whose observation densities are given in terms of linearly decomposable profile languages can be checked for equivalence in *polynomial time*, by a reduction to the finite-observation case. We also indicate, in example 3.6.1, how our result can be used to check for susceptibility of certain timing attacks.

Definition 3.0.1. For two distributions π_1 and π_2 and a HMM $\mathcal{C} = (Q, \Sigma, \Psi)$, we say that π_1 and π_2 are *equivalent*, written $\pi_1 \equiv_{\mathcal{C}} \pi_2$, if $\mathbb{P}_{\pi_1}(A) = \mathbb{P}_{\pi_2}(A)$ holds for all measurable subsets $A \subseteq \Sigma^\omega$.

One could define equivalence of two pairs $(\mathcal{C}_1, \pi_1)$ and $(\mathcal{C}_2, \pi_2)$ where $\mathcal{C}_i = (Q_i, \Sigma, \Psi_i)$ are HMMs and π_i are initial distributions for $i = 1, 2$. We do not need that though, as we can define, in a natural way, a single HMM over the disjoint union of Q_1 and Q_2 and consider instead equivalence of π_1 and π_2 (where π_1, π_2 are appropriately padded with zeros).

Given an observation density matrix Ψ, a *functional decomposition* consists of functions $f_k : \Sigma \to [0, \infty)$ and matrices $P_k \in \mathbb{R}^{|Q| \times |Q|}$ for $k \in [d]$ such that $\Psi(x) = \sum_{k=1}^d f_k(x) P_k$ for all $x \in \Sigma$ and $\int_\Sigma f_k \, d\lambda = 1$ for all $k \in [d]$. We sometimes abbreviate this decomposition as $\Psi = \sum_{k=1}^d f_k P_k$ and this notion has a central role in our thesis.

Example 3.0.1. The observation density matrix Ψ from example 2.1.1 has a functional decomposition

$$\begin{aligned}\Psi(x) &= 2\exp(-2x)\chi_{[0,\infty)}(x)\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix} + \chi_{[-1,0)}(x)\begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix} + \\ &\quad \frac{1}{2}\chi_{[0,2)}(x)\begin{pmatrix} 0 & 0 \\ \frac{1}{3} & 0 \end{pmatrix} + \exp(-x)\chi_{[0,\infty)}(x)\begin{pmatrix} 0 & 0 \\ 0 & \frac{2}{3} \end{pmatrix}. \end{aligned}$$ □

Lemma 3.0.2. *Let (Q, Σ, Ψ) be a HMM. If Ψ has functional decomposition $\Psi = \sum_{k=1}^d f_k P_k$ then $\sum_{k=1}^d P_k$ is stochastic.*

Proof. By definition of a HMM, $\int_\Sigma \Psi \, d\lambda$ is stochastic, and we have

$$\int_\Sigma \Psi \, d\lambda = \int_\Sigma \sum_{k=1}^d f_k P_k \, d\lambda = \sum_{k=1}^d P_k \int_\Sigma f_k \, d\lambda = \sum_{k=1}^d P_k.$$ □

When Σ is finite, it follows that $\int_\Sigma \Psi \, d\lambda = \sum_{a \in \Sigma} \Psi(a)$. Hence $\sum_{a \in \Sigma} \Psi(a)$ is stochastic.

Encoding For computational purposes we assume that rational numbers are represented as ratios of integers in binary. The initial distribution of a HMM with state set Q is given as a vector $\pi \in \mathbb{Q}^{|Q|}$. We also need to encode continuous functions, in particular, density functions such as Gaussian, exponential or piecewise-polynomial functions. A *profile* is a finite word (i.e., string) that describes a function. Its letters may be symbols or rational numbers that encode a function type and its parameters. For example, the profile $(\mathcal{N}, \mu, \sigma)$ may denote a Gaussian (also called normal) distribution with mean $\mu \in \mathbb{Q}$ and standard deviation $\sigma \in \mathbb{Q}_+$. A profile may also consist of a description of a rational linear combination of such building blocks. For any profile γ we write $[\![\gamma]\!] : \Sigma \to [0, \infty)$ for the function it encodes. For example, a profile $\gamma = (\mathcal{N}, \mu, \sigma)$ with $\mu \in \mathbb{Q},\ \sigma \in \mathbb{Q}_+$ may encode the function $[\![\gamma]\!] : \mathbb{R} \to [0, \infty)$ given as $[\![\gamma]\!](x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$. Without restricting ourselves to any particular encoding, we assume that Γ is a *profile language*, i.e., a finitely presented but usually infinite set of valid profiles. For any $\Gamma_0 \subseteq \Gamma$ we write $[\![\Gamma_0]\!] = \{[\![\gamma]\!] \mid \gamma \in \Gamma_0\}$.

We use profiles to encode HMMs $\mathcal{C} = (Q, \Sigma, \Psi)$: we say that $\mathcal{C}$ is *over* Γ if the observation density matrix Ψ is given as a matrix of pairs $(p_{i,j}, \gamma_{i,j}) \in \mathbb{Q}_+ \times \Gamma$ such that $\Psi_{i,j} = p_{i,j}[\![\gamma_{i,j}]\!]$ and $\int_\Sigma [\![\gamma_{i,j}]\!]\, d\lambda = 1$ hold for all $i, j \in [Q]$. In this way the $p_{i,j}$ form the transition probabilities between states and the $\gamma_{i,j}$ encode the probability densities of the observations upon each transition.

Example 3.0.3. For a suitable profile language Γ, the HMM from example 2.1.1 may be over Γ, with the observation density matrix given as

$$\begin{pmatrix} (\frac{1}{2}, (Exp, 2)) & (\frac{1}{2}, (U, -1, 0)) \\ (\frac{1}{3}, (U, 0, 2)) & (\frac{2}{3}, (Exp, 1)) \end{pmatrix}. \qquad \square$$

The observation density matrix Ψ of a HMM (Q, Σ, Ψ) with *finite* Σ can be given as a list of matrices $\Psi(a) \in \mathbb{Q}_+^{|Q|\times|Q|}$ for all $a \in \Sigma$ such that $\sum_{a\in\Sigma} \Psi(a)$ is a stochastic matrix.

3.1 Equivalence as Orthogonality

For finite-observation HMMs it is well known [12–15] that two initial distributions given as vectors $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ are equivalent if and only if $\pi_1 - \pi_2$ is orthogonal (written as $\perp$) to a certain vector space.

Proposition 3.1.1. *Consider a finite-observation HMM (Q, Σ, Ψ). For any $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ we have*

$$\pi_1 \equiv \pi_2 \iff \pi_1 - \pi_2 \perp \operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \Sigma^*\}.$$

In the finite-observation case, proposition 3.1.1 leads to an efficient algorithm for deciding equivalence. We will sometimes refer to $\operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \Sigma^*\}$ as the *state distribution span.*

Proposition 3.1.2. *Let (Q, Σ, Ψ) be a finite-observation HMM. One can compute a basis $\mathcal{B}$ for $\operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \Sigma^*\}$ in polynomial time.*

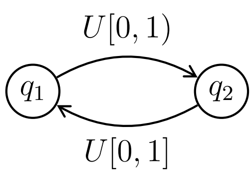
We provide a sketch proof. More detail can be found in [12, 14].

Sketch Proof. This can be done using a fixed-point algorithm that computes a sequence of (bases of) increasing subspaces of $\mathcal{V} := \operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \Sigma^*\}$: start with $\mathcal{B} = \{\overline{1}^\top\}$, and as long as $|\mathcal{B}| < |Q|$ or there is $a \in \Sigma$ and $v \in \mathcal{B}$ such that $\Psi(a)v \notin \operatorname{span} \mathcal{B}$, add $\Psi(a)v$ to $\mathcal{B}$. Since $\dim \mathcal{V} \leq |Q|$, this algorithm terminates after at most $|Q|$ iterations, and returns $\mathcal{B}$ such that $\operatorname{span} \mathcal{B} = \mathcal{V}$. □

It is then easy to check whether $\pi_1 - \pi_2 \perp \mathcal{V}$. It follows:

Proposition 3.1.3. *Given a HMM (Q, Σ, Ψ) with finite Σ and initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^{|Q|}$, it is decidable in polynomial time whether $\pi_1 \equiv \pi_2$.*

Example 3.1.4. In general Proposition 3.1.2 does not hold in the continuous-observation case. Consider the HMM:



Here, $U[0,1)$ and $U[0,1]$ are uniform distributions on $[0,1)$ and $[0,1]$ respectively. Assume that $\Sigma = [0,1]$ and has measure space $(\Sigma, \mathscr{G}, \lambda_{Leb})$. We have

$$\Psi(x) = \begin{cases} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & x \in [0,1) \\ \\ \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} & x = 1 \end{cases} \tag{3.1}$$

and so $\operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \Sigma^*\} = \operatorname{span}\{(1,1)^\top, (0,1)^\top\} = \mathbb{R}^2$. It follows that $\pi_1 - \pi_2 \perp \operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \Sigma^*\}$ if and only if $\pi_1 = \pi_2$.

On the other hand consider $\pi_1 = e_{q_1}$ and $\pi_2 = e_{q_2}$. Let $E_1 \times \cdots \times E_n \in \mathscr{G}^n$ be a cylinder set. For each $i \in [n]$ we define $E_i' = E_i \setminus \{1\}$. We have $\lambda_{Leb}(E_1 \times \cdots \times E_n) = \lambda_{Leb}(E_1' \times \cdots \times E_n')$. Hence for $i = 1, 2$:

$$\begin{aligned}
\mathbb{P}_{e_{q_i}}(E_1 \times \cdots \times E_n \Sigma^\omega) &= e_{q_i} \int_{E_1 \times \cdots \times E_n} \Psi d\lambda^n \overline{1}^\top \\
&= e_{q_i} \int_{E_1' \times \cdots \times E_n'} \Psi d\lambda^n \overline{1}^\top \\
&= \int_{E_1' \times \cdots \times E_n'} e_{q_i} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \overline{1}^\top d\lambda^n \\
&= \lambda^n(E_1' \times \cdots \times E_n').
\end{aligned}$$

It follows that $\mathbb{P}_{e_{q_1}}(E_1 \times \cdots \times E_n \Sigma^\omega) = \mathbb{P}_{e_{q_2}}(E_1 \times \cdots \times E_n \Sigma^\omega)$ and so $e_{q_1} \equiv e_{q_2}$, but $e_{q_1} \neq e_{q_2}$.

The observation $1 \in \Sigma$ occurs at a discontinuity of Ψ and only gets produced on a λ_{Leb}-null set. However, it still gets represented in $\operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \Sigma^*\}$. By excluding the discontinuities of Ψ from Σ, we show that a version of Proposition 3.1.2 holds for continuous-observation HMMs.

As discussed in the preliminaries, Ψ is almost-surely continuous with set of continuity $\mathring{\Sigma} \subseteq \Sigma$ such that $\lambda(\mathring{\Sigma}) = \lambda(\Sigma)$ and Ψ is continuous on $\mathring{\Sigma}$. We prove the following in Section 3.4.

Proposition 3.1.5. *Consider a continuous-observation HMM (Q, Σ, Ψ). For any $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ we have*

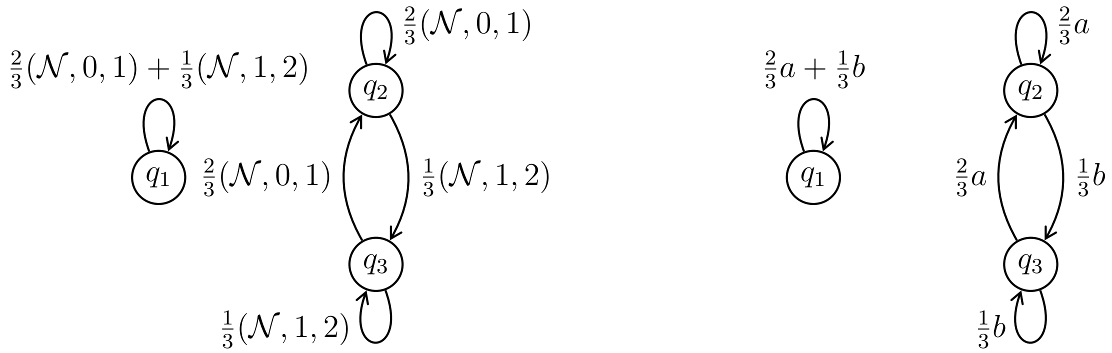
$$\pi_1 \equiv \pi_2 \iff \pi_1 - \pi_2 \perp \operatorname{span}\left\{\Psi(w)\overline{1}^\top \,\middle|\, w \in \mathring{\Sigma}^*\right\}.$$

The algorithm used in Proposition 3.1.2 is not an effective algorithm when Σ is infinite.

3.2 Labelling Reductions

Our goal is to reduce in polynomial time the equivalence problem in continuous-observation HMMs to the equivalence problem in finite-observation HMMs. Since the latter is decidable in polynomial time by proposition 3.1.3, a polynomial time algorithm for deciding equivalence in continuous-observation HMMs follows.

Towards this objective, consider a reduction where each continuous density function is given a label and these labels form the observation alphabet of a finite-observation HMM. For example consider the HMM on the left in the diagram below. This disconnected HMM emits letters from two distinct normal distributions with profiles $(\mathcal{N}, 0, 1)$ and $(\mathcal{N}, 1, 2)$. Assigning each distribution letters a, b respectively yields the HMM given on the right. Since in the right HMM states q_1 and q_2 are equivalent so too are the same labelled states in the continuous HMM.



More rigorously, if $\mathcal{C} = (Q, \Sigma, \Psi)$ is a HMM over $\Gamma = \{\beta_1, \ldots, \beta_K\}$ and Ψ is encoded as a matrix of coefficient-profile pairs $(p_{i,j}, \gamma_{i,j}) \in \mathbb{Q}_+ \times \Gamma$ then we call the *labelling reduction* the HMM $(Q, \hat{\Sigma}, \hat{M})$ where $\hat{\Sigma} = \{a_1, \ldots, a_K\}$ is an alphabet of fresh observations and

$$\hat{M}_{i,j}(a_k) = \begin{cases} p_{i,j} & \gamma_{i,j} = \beta_k \\ 0 & \text{otherwise.} \end{cases}$$

Since Ψ has functional decomposition $\Psi = \sum_{k=1}^{K} [\![\beta_k]\!] \hat{M}(a_k)$, it follows by Lemma 3.0.2 that $\sum_{k=1}^{K} \hat{M}(a_k)$ is stochastic and the labelling reduction is a well defined HMM which can be computed in polynomial time. As discussed in the previous example, equivalence in the labelling reduction implies equivalence in the original HMM:

Proposition 3.2.1. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a HMM with labelling reduction $L = (Q, \hat{\Sigma}, \hat{M})$. Then for any initial distributions π_1 and π_2*

$$\pi_1 \equiv_L \pi_2 \implies \pi_1 \equiv_{\mathcal{C}} \pi_2.$$

For the proof of Proposition 3.2.1 we use the following lemma which will be re-used in section 3.5.

Lemma 3.2.2. *Let* $\mathcal{C}_1 = (Q, \Sigma_1, \Psi_1)$ *and* $\mathcal{C}_2 = (Q, \Sigma_2, \Psi_2)$ *be two HMMs with the same state space* Q*. Suppose that* $\mathrm{span}\{\Psi_1(x) \mid x \in \Sigma_1\} \subseteq \mathrm{span}\{\Psi_2(x) \mid x \in \Sigma_2\}$*. Then,* $\mathrm{span}\{\Psi_1(w) \mid w \in \Sigma_1^*\} \subseteq \mathrm{span}\{\Psi_2(w) \mid w \in \Sigma_2^*\}$ *and for any two initial distributions* π_1 *and* π_2*,*

$$\pi_1 \equiv_{\mathcal{C}_2} \pi_2 \implies \pi_1 \equiv_{\mathcal{C}_1} \pi_2.$$

Proof. Let $w = x_1 \cdots x_N \in \Sigma_1^*$. Then $\Psi_1(x_n) = \sum_{i=1}^{I_n} \lambda_{i,n} \Psi_2(y_{i,n})$ for $n \in [N]$ and

$$\begin{aligned}
\Psi_1(w) &= \Big(\sum_{i_1=1}^{I_1} \lambda_{i_1,1}\Psi_2(y_{i_1,1})\Big) \cdots \Big(\sum_{i_N=1}^{I_N} \lambda_{i_N,N}\Psi_2(y_{i_N,N})\Big) \\
&= \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \lambda_{i_1,1} \ldots \lambda_{i_N,N} \Psi_2(y_{i_1,1}) \ldots \Psi_2(y_{i_N,N}) \;\in\; \mathrm{span}\{\Psi_2(w) \mid w \in \Sigma_2^*\}.
\end{aligned}$$

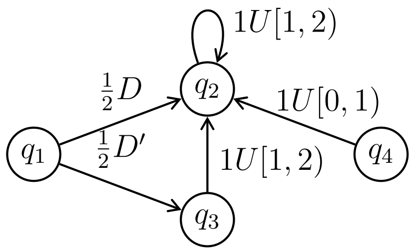
Thus, $\mathrm{span}\{\Psi_1(w) \mid w \in \Sigma_1^*\} \subseteq \mathrm{span}\{\Psi_2(w) \mid w \in \Sigma_2^*\}$. Therefore, by Proposition 3.1.5,

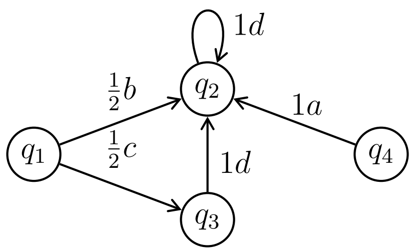
$$\begin{aligned}
\pi_1 \equiv_{\mathcal{C}_2} \pi_2 &\iff \pi_1 - \pi_2 \perp \mathrm{span}\{\Psi_2(w)\overline{1}^\top \mid w \in \Sigma_2^*\} \\
&\implies \pi_1 - \pi_2 \perp \mathrm{span}\{\Psi_1(w)\overline{1}^\top \mid w \in \Sigma_1^*\} \\
&\iff \pi_1 \equiv_{\mathcal{C}_1} \pi_2.
\end{aligned}$$ □

Proof of proposition 3.2.1. Ψ has a functional decomposition $\Psi = \sum_{k=1}^K [\![\beta_k]\!] \hat{M}(a_k)$. Thus, $\mathrm{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\} \subseteq \mathrm{span}\{\hat{M}(a_k) \mid a_k \in \hat{\Sigma}\}$ and the statement follows by lemma 3.2.2. □

Example 3.2.3. Consider the HMMs in the diagram below and recall that for a set $A \subseteq \Sigma$, $\chi_A : \Sigma \to \{0,1\}$ is the standard indicator function. The HMM on the left is a continuous-observation HMM where D and D' are distributions on $[0,1]$ with probability density functions $2x\chi_{[0,1)}(x)$ and $2(1-x)\chi_{[0,1)}(x)$ respectively, and $U[a,b)$ is the uniform distribution on $[a,b)$. The HMM on the right is the corresponding labelling reduction.

Since $U[0,1) = \frac{1}{2}D + \frac{1}{2}D'$, (the Dirac distributions on) states q_1 and q_4 are equivalent but as the distributions $U[0,1), D, D'$ are distinct, they get assigned different labels a, b, c, respectively in the labelling reduction. The states q_1 and q_4 are therefore not equivalent in the right HMM.





3.3 Linearly Decomposable Profile Languages

Example 3.2.3 shows that the linear combination of two continuous distributions can "imitate" a single distribution. Therefore we consider the transition densities as part of a vector space of functions. In the usual way $\mathcal{L}_1(\Sigma, \lambda)$ is the quotient vector space where functions that differ only on a λ-null set are identified. For $f, g : \Sigma \to \mathbb{R}$ we say $f =_\lambda g$ if $\{x \in \Sigma \mid f(x) \neq g(x)\}$ is a λ-null set. In particular, when $\Sigma \subseteq \mathbb{R}$ and λ is the Lebesgue measure λ_{Leb}, we have $\chi_{[a,b)} =_{\lambda_{Leb}} \chi_{(a,b]}$.

Definition 3.3.1. Let Γ be a profile language with $[\![\Gamma]\!] \subseteq \mathcal{L}_1(\Sigma, \lambda)$. We say that Γ is *linearly decomposable* if there is a procedure that takes a finite subset $\{\gamma_1, \ldots, \gamma_n\} = \Gamma_0 \subseteq \Gamma$ and produces profiles $\beta_1, \ldots, \beta_m \in \Gamma_0$ and a set of rational coefficients $b_{i,j} \in \mathbb{Q}$ for $i \in [n], j \in [m]$ such that $\{[\![\beta_1]\!], \ldots, [\![\beta_m]\!]\}$ is a basis for $\mathrm{span}\{[\![\gamma_1]\!], \ldots, [\![\gamma_n]\!]\}$ (hence $m \leq n$), and

$$[\![\gamma_i]\!] =_\lambda \sum_{j=1}^{m} b_{i,j} [\![\beta_j]\!] \text{ for all } i \in [n]. \tag{3.2}$$

Further, this procedure takes polynomial time in the total length of the encoding of Γ_0.

The following theorem is the main result of this chapter:

Theorem 3.3.1. *Given a HMM $\mathcal{C} = (Q, \Sigma, \Psi)$ over a linearly decomposable profile language, and initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^{|Q|}$, it is decidable in polynomial time (in the size of the encoding) whether $\pi_1 \equiv \pi_2$.*

Like in the finite case, to prove theorem 3.3.1 we use the following continuous analogue of Proposition 3.1.2 for computing the state distribution span. This result is also used later in Chapter 4.

Proposition 3.3.2. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a continuous-observation HMM over a linearly decomposable profile language. Suppose Ψ has set of continuity $\mathring{\Sigma}$. We may compute a basis $\mathcal{B}$ for* $\operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\}$ *in polynomial time (in the size of the encoding of $\mathcal{C}$).*

We prove proposition 3.3.2 in section 3.5.

Proof of Theorem 3.3.1. By Proposition 3.3.2 we may compute a basis $\mathcal{B}$ for $\operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\}$ in polynomial time. We have

$$\begin{aligned} \pi_1 \equiv_{\mathcal{C}} \pi_2 \iff & \pi_1 - \pi_2 \perp \operatorname{span}\left\{\Psi(w)\bar{1}^\top \;\middle|\; w \in \mathring{\Sigma}^*\right\} \text{ by Proposition 3.1.5} \\ \iff & (\pi_1 - \pi_2) \cdot b \quad \forall b \in \mathcal{B}. \end{aligned} \tag{3.3}$$

The condition (3.3) can be checked in polynomial time. $\square$

To make the notion of linearly decomposable profile languages more concrete, we give a concrete example in the following subsection.

Example: Gaussian, Exponential, and Piecewise Polynomial Functions

We describe a profile language, Γ_{GEM}, that can specify linear combinations Gaussian, exponential, and piecewise polynomial density functions.

Let $\mathbb{I} = \{(a, b) \subseteq \mathbb{R} \mid a, b \in \mathbb{Q} \cup \{-\infty, \infty\} : a < b\}$ and let γ be a profile. We define a new profile $\mathcal{I}(\gamma, a, b)$ obtained by restricting the support of $[\![\gamma]\!]$ to an interval $(a, b) \in \mathbb{I}$.

We call a function of the form $x \mapsto x^k \chi_I(x)$ where $k \in \mathbb{N}_0$ and $I \in \mathbb{I}$ is an interval an *interval-domain monomial.* To avoid clutter, we often denote interval-domain monomials only by $x^k \chi_I$. Any piecewise polynomial is a linear combination of interval-domain monomials.

Suppose γ encodes the monomial function $x \mapsto x^k$ for $k \in \mathbb{N}_0$. Then, $\mathcal{I}(\gamma, a, b)$ encodes the interval-domain monomial $x \mapsto x^k \chi_{(a,b)}(x)$ where $a, b \in \mathbb{Q} \cup \{-\infty, \infty\}$ and $a < b$.

Let M be a set of profiles encoding monomials x^k in terms of $k \in \mathbb{N}_0$. We then set $M' = \{\mathcal{I}(\gamma, a, b) \mid \gamma \in M, (a, b) \in \mathbb{I}\}$ which is a set of profiles that can encode any interval-domain monomial function.

Gaussian and exponential density functions can be fully described using their parameters, which we assume to be rational. We write G and E for corresponding sets of profiles, respectively. Finally, we fix a profile language $\Gamma_{GEM} \supset G \cup E \cup M'$ obtained by closing $G \cup E \cup M'$ under linear combinations. That is, for any $\gamma_1, \ldots, \gamma_k \in \Gamma_{GEM}$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{Q}$, there exists a profile $\gamma \in \Gamma_{GEM}$ such that $[\![\gamma]\!] = \lambda_1 [\![\gamma_1]\!] + \cdots + \lambda_k [\![\gamma_k]\!]$. This closure can be achieved using a specific constructor, say $\mathcal{S}$, for linear combinations, so that $\gamma = \mathcal{S}(\lambda_1, \gamma_1, \ldots, \lambda_k, \gamma_k)$. We have:

$$\Gamma_{GEM} = \{\mathcal{S}(\lambda_1, \gamma_1, \ldots, \lambda_k, \gamma_k) \mid \gamma_i \in G \cup E \cup M',\ \lambda_i \in \mathbb{Q},\ i \in [k]\}\,.$$

Example 3.3.3. The continuous-observation HMM $(Q, \mathbb{R}, \Psi)$ from Example 3.2.3 is over Γ_{GEM}: the observation density matrix Ψ can be encoded as a matrix of coefficient-profile pairs

$$\begin{pmatrix} 0 & (\frac{1}{2}, \gamma_1) & (\frac{1}{2}, \gamma_2) & 0 \\ 0 & (1, \gamma_3) & 0 & 0 \\ 0 & (1, \gamma_3) & 0 & 0 \\ 0 & (1, \gamma_4) & 0 & 0 \end{pmatrix}$$

with $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \Gamma_{GEM}$ and $[\![\gamma_1]\!] = 2x\chi_{[0,1)}$ and $[\![\gamma_2]\!] = 2(1-x)\chi_{[0,1)}$ and $[\![\gamma_3]\!] = \chi_{[1,2)}$ and $[\![\gamma_4]\!] = \chi_{[0,1)}$. $\square$

Let H be a set of disjoint half open intervals. Suppose that $m_1, \ldots, m_I$ are distinct interval-domain monomials such that $\mathrm{supp}(m_i) \in H$ for all $i \in [I]$. In addition, let $g_1, \ldots, g_J$ and $e_1, \ldots, e_K$ be distinct Gaussian and exponential density functions, respectively. We show later in Chapter 5 that the set

$$\{m_1, \ldots, m_I, g_1, \ldots, g_J, e_1, \ldots, e_K\}$$

is linearly independent. Therefore, by the construction of Γ_{GEM}, we can compute the profiles and coefficients in (3.2) from the coefficients in the description of Ψ as a matrix of profile-coefficient pairs. We also prove the following in Chapter 5.

Proposition 3.3.4. *The profile language Γ_{GEM} is linearly decomposable.*

Thus we obtain the following corollary of theorem 3.3.1:

Corollary 3.3.5. *Given a HMM (Q, Σ, Ψ) over Γ_{GEM}, and initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^{|Q|}$, it is decidable in polynomial time whether $\pi_1 \equiv \pi_2$.*

3.4 The Set of Continuity

Recall that an observation density matrix $\Sigma : \Sigma \to [0,\infty)^{|Q|\times|Q|}$ is almost-surely continuous and so has set of continuity $\mathring{\Sigma}$. This section is dedicated to the proof of Proposition 3.1.5. We restate it here.

Proposition 3.1.5. *Consider a continuous-observation HMM* (Q,Σ,Ψ)*. For any* $\pi_1,\pi_2 \in \mathbb{R}^{|Q|}$ *we have*

$$\pi_1 \equiv \pi_2 \iff \pi_1 - \pi_2 \perp \operatorname{span}\left\{\Psi(w)\overline{1}^\top \;\middle|\; w \in \mathring{\Sigma}^*\right\}.$$

We use the following intermediate lemma.

Lemma 3.4.1. *Let* (Q,Σ,Ψ) *be a HMM and let* π_1,π_2 *be initial distributions. Fix* $n \in \mathbb{N}$*. Then,* $(\pi_1-\pi_2)\Psi(w)\overline{1}^\top = 0$ *for all* $w \in \mathring{\Sigma}^n$ *if and only if* $(\pi_1 - \pi_2)\int_E \Psi \, d\lambda^n \overline{1}^\top = 0$ *for all* $E \in \mathscr{G}^n$*.*

Proof. For the forward implication, suppose $E \in \mathscr{G}^n$ then $E \setminus (E \cap \mathring{\Sigma})$ is a λ-null set. Thus,

$$(\pi_1-\pi_2)\int_E \Psi \, d\lambda^n \overline{1}^\top = (\pi_1-\pi_2)\int_{E\cap\mathring{\Sigma}} \Psi \, d\lambda^n \overline{1}^\top = 0.$$

For the converse implication, suppose there is $v \in \mathring{\Sigma}^n$ such that $(\pi_1-\pi_2)\Psi(v)\overline{1}^\top > 0$. Since Ψ is continuous when restricted to $\mathring{\Sigma}^n$, there is $\varepsilon > 0$ and an open ball $B(v,\varepsilon)$ such that $(\pi_1-\pi_2)\Psi(w)\overline{1}^\top > 0$ for all $w \in B(v,\varepsilon)$. It follows that $(\pi_1-\pi_2)\int_{B(v,\varepsilon)} \Psi \, d\lambda^n \overline{1}^\top > 0$. A symmetrical argument can be applied in the case $(\pi_1-\pi_2)\Psi(v)\overline{1}^\top < 0$. ☐

We may now prove Proposition 3.1.5.

Proof of Proposition 3.1.5. We have:

$$\begin{aligned}
\pi_1 \equiv \pi_2 &\iff \mathbb{P}_{\pi_1}(E) = \mathbb{P}_{\pi_2}(E) && \forall E \in \mathscr{G}^* \\
&\iff (\pi_1-\pi_2)\int_E \Psi \, d\lambda^n \overline{1}^\top = 0 && \forall E \in \mathscr{G}^n,\ n \in \mathbb{N} \\
&\iff (\pi_1-\pi_2)\Psi(w)\overline{1}^\top = 0 && \forall w \in \mathring{\Sigma}^n,\ n \in \mathbb{N} \\
&\iff (\pi_1-\pi_2) \perp \operatorname{span}\{\Psi(w)\overline{1}^\top \mid w \in \mathring{\Sigma}^*\},
\end{aligned}$$

where the third equivalence follows from Lemma 3.4.1 and is a result of Ψ being continuous on $\mathring{\Sigma}$. ☐

3.5 Independent Functional Decompositions

This section is dedicated to the proof of Proposition 3.3.2. Suppose that Ψ has a functional decomposition $\sum_{k=1}^{d} f_k P_k$ such that the set $\{f_1, \ldots, f_d\}$ is linearly independent. Then, $\sum_{k=1}^{d} f_k P_k$ is called an *independent functional decomposition*. The efficient computation of an independent functional decomposition is the key ingredient for the proof of theorem 3.3.1. We start with the following lemma.

Lemma 3.5.1. *Suppose* $\Psi : \Sigma \to [0,\infty)^{|Q|\times|Q|}$ *has an independent functional decomposition* $\Psi = \sum_{k=1}^{d} f_k P_k$. *Then,* $\mathrm{span}\{\Psi(x) \mid x \in \Sigma\} = \mathrm{span}\{P_k \mid k \in [d]\}$.

For the proof of this lemma we need a result concerning *alternant matrices*. Consider functions $f_1, \ldots, f_n : \Sigma \to \mathbb{R}$ and let $x_1, \ldots, x_n \in \Sigma$. Then,

$$M = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_n) & f_2(x_n) & \ldots & f_n(x_n) \end{pmatrix}$$

is called the alternant matrix for $f_1, \ldots, f_n$ and *input points* $x_1, \ldots, x_n$. The following is proven in Appendix A.1.

Lemma 3.5.2. *Suppose* $f_1, \ldots, f_n \in \mathcal{L}_1(\Sigma, \lambda)$. *Then, the* f_i *are linearly dependent if and only if all alternant matrices for the* f_i *are singular.*

Proof of Lemma 3.5.1. Since $\Psi(x) = \sum_{k=1}^{d} f_k(x) P_k$, we have $\mathrm{span}\{\Psi(x) \mid x \in \Sigma\} \subseteq \mathrm{span}\{P_k \mid k \in [d]\}$. For the reverse inclusion, since the f_i are linearly independent, by Lemma 3.5.2 there exists an alternant matrix M with full rank for $f_1, \ldots, f_d$ with input points $x_1, \ldots, x_d$. Hence, for each of the standard basis vectors $e_k \in \{0,1\}^d$, $k \in [d]$, there exists $v_k = (v_{1,k}, \ldots, v_{d,k}) \in \mathbb{R}^d$ such that $v_k M = e_k$. Writing $\delta_{j,k}$ for the Kronecker delta function it follows that

$$\sum_{i=1}^{d} v_{i,k} \Psi(x_i) \;=\; \sum_{i=1}^{d} v_{i,k} \sum_{j=1}^{d} f_j(x_i) P_j \;=\; \sum_{j=1}^{d} P_j \sum_{i=1}^{d} v_{i,k} f_j(x_i) \;=\; \sum_{j=1}^{d} P_j \delta_{j,k} \;=\; P_k \,,$$

which implies that $\mathrm{span}\{\Psi(x) \mid x \in \Sigma\} \supseteq \mathrm{span}\{P_k \mid k \in [d]\}$. $\square$

The proof of the following proposition re-uses lemma 3.2.2 from section 3.2.

Proposition 3.5.3. *Suppose that HMM $\mathcal{C} = (Q, \Sigma, \Psi)$ has independent functional decomposition $\Psi = \sum_{k=1}^{d} f_k P_k$ and each P_k is non-negative for all $k \in [d]$. Define a set $\overline{\Sigma} = \{a_1, \ldots, a_d\}$ of fresh observations and the observation density matrix M with $M(a_k) = P_k$ for all $k \in [d]$. Then $\mathcal{F} = (Q, \overline{\Sigma}, M)$ is a finite-observation HMM and for any initial distributions π_1, π_2*

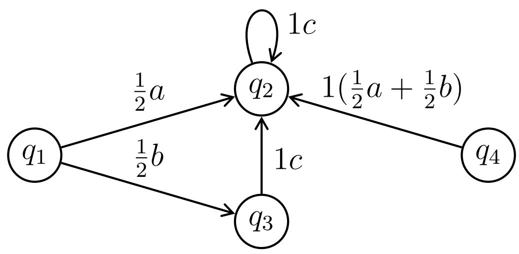
$$\pi_1 \equiv_{\mathcal{C}} \pi_2 \iff \pi_1 \equiv_{\mathcal{F}} \pi_2.$$

Proof. It follows by Lemma 3.0.2 that $\sum_{k=1}^{d} P_k$ is stochastic. Thus F defines a HMM. By Lemma 3.5.1, $\text{span}\{\Psi(x)\overline{1}^{\top} \mid x \in \Sigma\} = \text{span}\{M(a)\overline{1}^{\top} \mid a \in \overline{\Sigma}\}$ which combined with Lemma 3.2.2 gives the result. □

Example 3.5.4. We use the HMM $\mathcal{C}$ discussed in Examples 3.2.3 and 3.3.3 to illustrate the construction of Proposition 3.5.3. The basis $\{2x\chi_{[0,1)}, 2(1-x)\chi_{[0,1)}, \chi_{[1,2)}\}$ leads to the independent functional decomposition

$$\Psi = 2x\chi_{[0,1)} \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} + 2(1-x)\chi_{[0,1)} \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} + \chi_{[1,2)} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Therefore, Proposition 3.5.3 implies that two initial distributions $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ are equivalent in $\mathcal{C}$ if and only if they are equivalent in the following HMM:



Here, states q_1 and q_4 are equivalent. Hence, they are also equivalent in $\mathcal{C}$. □

If an observation density matrix has an entry with pdf $2e^{-x} - 2e^{-2x}$ (which is encodable in Γ_{GEM} due to its linear combination closure property), the independent functional decomposition generated by the algorithm described in the proof of Proposition 3.3.4 has matrices which are not all non-negative. Therefore, Proposition 3.5.3 cannot be applied directly. However, given an independent functional decomposition $\Psi = \sum_{k=1}^{d} f_k P_k$ and noting that $\sum_{k=1}^{d} P_k$ is stochastic by Lemma 3.0.2, the following proposition shows that there is a small $\theta > 0$ such that $P - \theta P_k$ is non-negative for all $k \in [d]$. Furthermore, $\mathrm{span}\{P_k \mid k \in [d]\} = \mathrm{span}\{P - \theta P_k \mid k \in [d]\}$. These two facts lead us to construct a finite-observation HMM using the scaled transition matrices $\frac{1}{d-\theta}(P - \theta P_k)$.

Proposition 3.5.5. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a HMM with independent functional decomposition $\Psi = \sum_{k=1}^{d} f_k P_k$. Let $P = \sum_{k=1}^{d} P_k$ and*

$$\theta = \min\left\{\frac{1}{2}, \frac{\min\{(P)_{i,j} \mid (P)_{i,j} > 0\}}{\max\{\big(P_k\big)_{i,j} \mid i,j \in [Q],\ k \in [d]\}}\right\}.$$

Define an alphabet $\tilde{\Sigma} = \{a_1, \ldots, a_d\}$ of fresh observations and the HMM $\mathcal{F} = (Q, \tilde{\Sigma}, M)$ with $M(a_k) = \frac{1}{d-\theta}(P - \theta P_k)$. Then

$$\mathrm{span}\{M(a_k)\bar{1}^\top \mid k \in [d]\} = \mathrm{span}\{\Psi(x)\bar{1}^\top \mid x \in \Sigma\}$$

and for any initial distributions π_1, π_2,

$$\pi_1 \equiv_{\mathcal{C}} \pi_2 \iff \pi_1 \equiv_{\mathcal{F}} \pi_2.$$

Proof. First we show that $\mathcal{F}$ is a well-defined HMM. Matrix $\sum_{k=1}^{d} M(a_k)$ is stochastic as

$$\sum_{k=1}^{d} M(a_k) = \frac{1}{d-\theta}\sum_{k=1}^{d}(P - \theta P_k) = \frac{dP - \theta\sum_{k=1}^{d} P_k}{d-\theta} = P\,, \tag{3.4}$$

and by Lemma 3.0.2, P is stochastic. In addition we must show that $M(a_k)$ is non-negative for each $k \in [d]$. Since $\theta \le \frac{1}{2}$, it is enough to show that $P - \theta P_k$ is non-negative for each $k \in [d]$. Suppose that $(P)_{i,j} = 0$. Then, $\int_\Sigma \Psi_{i,j}\, d\lambda = (P)_{i,j} = 0$, which implies that $\Psi_{i,j} = 0$ since Ψ is piecewise continuous. Thus, $\sum_{k=1}^{d} f_k (P_k)_{i,j} = \Psi_{i,j} = 0$. Since $\{f_k\}_{k=1}^{d}$ is linearly independent, it follows that $(P_k)_{i,j} = 0$ for all $k \in [d]$ and so $(P - \theta P_k)_{i,j} = 0$. Now suppose that $(P)_{i,j} > 0$. By the definition of θ, it follows that $(\theta P_k)_{i,j} \le (P)_{i,j}$. Thus, $\mathcal{F}$ is a well defined HMM.

Observe that $\mathrm{span}\{P - \theta P_k \mid k \in [d]\} \subseteq \mathrm{span}\{P_k \mid k \in [d]\}$. The opposite inclusion follows from the fact that, by Equation (3.4), we have

$$P \in \mathrm{span}\{P - \theta P_k \mid k = 1, \ldots, d\}.$$

Thus, by Lemma 3.5.1,

$$\begin{aligned}\operatorname{span}\{M(a) \mid a \in \tilde{\Sigma}\} &= \operatorname{span}\{P - \theta P_k \mid k \in [d]\} \\ &= \operatorname{span}\{P_k \mid k \in [d]\} \\ &= \operatorname{span}\{\Psi(x) \mid x \in \Sigma\}\,.\end{aligned}$$

The second point in the proposition follows immediately by Lemma 3.2.2. □

We may now prove the following:

Proposition 3.3.2. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a continuous-observation HMM over a linearly decomposable profile language. Suppose Ψ has set of continuity $\mathring{\Sigma}$. We may compute a basis $\mathcal{B}$ for $\operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\}$ in polynomial time (in the size of the encoding of $\mathcal{C}$).*

Proof. Suppose the HMM $\mathcal{C}$ is over the linearly decomposable profile language Γ. Let $\Gamma_0 = \{\gamma_1, \ldots, \gamma_n\}$ be the set of profiles appearing in the description of Ψ. From the description of Ψ as a matrix of coefficient-profile pairs, we can easily compute matrices $P'_1, \ldots, P'_n \in \mathbb{Q}^{|Q|\times|Q|}$ such that $\Psi = \sum_{i=1}^{n} [\![\gamma_i]\!] P'_i$. Since Γ is linearly decomposable, one can compute in polynomial time a subset $\{\beta_1, \ldots, \beta_d\} \subseteq \Gamma_0$ such that $[\![\{\beta_1, \ldots, \beta_d\}]\!]$ is linearly independent and also a set of coefficients $b_{i,k}$ such that $[\![\gamma_i]\!] =_\lambda \sum_{k=1}^{d} b_{i,k} [\![\beta_k]\!]$ for all $i \in [n]$. Hence:

$$\Psi \;=\; \sum_{i=1}^{n} [\![\gamma_i]\!] P'_i \;=_\lambda\; \sum_{i=1}^{n}\sum_{k=1}^{d} [\![\beta_k]\!] b_{i,k} P'_i \;=\; \sum_{k=1}^{d} [\![\beta_k]\!] \sum_{i=1}^{n} b_{i,k} P'_i$$

Setting $P_k = \sum_{i=1}^{n} b_{i,k} P'_i$ for all $k \in [d]$, we have $\Psi =_\lambda \sum_{k=1}^{d} [\![\beta_k]\!] P_k$.

Let $\Sigma_\lambda = \{x \in \Sigma \mid \Psi(x) = \sum_{k=1}^{d} [\![\beta_k]\!](x) P_k\}$. We define $\mathcal{C}' = (Q, \Sigma_\lambda \cap \mathring{\Sigma}, \Psi')$ where Ψ' is the restriction of Ψ to $\Sigma_\lambda \cap \mathring{\Sigma}$. We have $\mathcal{C}'$ is a valid HMM because Ψ' is stochastic:

$$\int_{\Sigma_\lambda \cap \mathring{\Sigma}} \Psi' d\lambda = \int_{\Sigma_\lambda \cap \mathring{\Sigma}} \Psi d\lambda = \int_{\Sigma} \Psi d\lambda \quad \text{since } \lambda(\Sigma \setminus (\Sigma_\lambda \cap \mathring{\Sigma})) = 0.$$

Further, $\Psi'(x) = \sum_{k=1}^{d} [\![\beta_k]\!](x) P_k$ for all $x \in \Sigma_\lambda \cap \mathring{\Sigma}$ and so the restriction of $\sum_{k=1}^{d} [\![\beta_k]\!] P_k$ to $\Sigma_\lambda \cap \mathring{\Sigma}$ is an independent functional decomposition of Ψ'. Therefore, from $P_1, \ldots, P_d$ we may compute $\mathcal{F} = (Q, \tilde{\Sigma}, M)$ from Proposition 3.5.5 and

$\mathrm{span}\{M(a)\overline{1}^\top \mid a \in \tilde{\Sigma}\} = \mathrm{span}\{\Psi(w)\overline{1}^\top \mid x \in \Sigma_\lambda \cap \mathring{\Sigma}\}$. By Lemma 3.2.2 we have

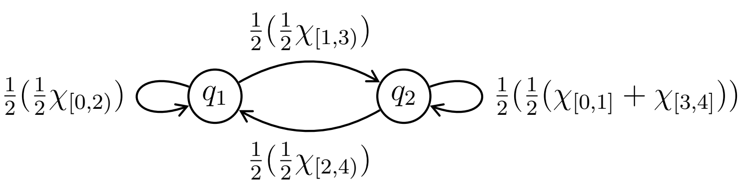
$$\begin{aligned}\mathrm{span}\{M(u)\overline{1}^\top \mid u \in \tilde{\Sigma}^*\} &= \mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in (\Sigma_\lambda \cap \mathring{\Sigma})^*\}\\ &\subseteq \mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in \mathring{\Sigma}^*\}. \qquad (3.5)\end{aligned}$$

It remains to show the final set inclusion in (3.5) is an equality. Fix $\mu \in \mathbb{R}^{|Q|}$. We show $\mu \perp \mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in (\Sigma_\lambda \cap \mathring{\Sigma})^*\}$ implies $\mu \perp \mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in \mathring{\Sigma}^*\}$.

Suppose $\mu\Psi(w)\overline{1}^\top > 0$ for some $w \in \mathring{\Sigma}^*$. If w is the empty word ε then $\varepsilon \in (\Sigma_\lambda \cap \mathring{\Sigma})^*$ and so $\mu \not\perp \mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in (\Sigma_\lambda \cap \mathring{\Sigma})^*\}$. In the case w has length $n \in \mathbb{N}$ then there is $\delta > 0$ and open ball $B(w,\delta)$ such that $\mu\Psi(w')\overline{1}^\top > 0$ for all $w' \in B(w,\delta) \cap \mathring{\Sigma}^{|w|}$. We have $B(w,\delta)$ is an open set and so has λ-positive measure by the assumption in the preliminaries. It follows that there is some $w' \in (\Sigma_\lambda \cap \mathring{\Sigma})^*$ such that $\mu\Psi(w')\overline{1}^\top > 0$. We use a symmetrical argument in the case $\mu\Psi(w)\overline{1}^\top < 0$.

With $\mathcal{F}$ computed, using Proposition 3.1.2 we may compute in polynomial time a basis $\mathcal{B}$ for $\mathrm{span}\{M(u)\overline{1}^\top \mid u \in \tilde{\Sigma}^*\}$. The basis $\mathcal{B}$ is therefore also a basis for $\mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in \mathring{\Sigma}^*\}$. □

Example 3.5.6. We illustrate aspects of the proof of Proposition 3.3.2 using the HMM:



Let Ψ' be the restriction of Ψ to $\mathring{\Sigma} = [0,1) \cup (1,2) \cup (2,3) \cup (3,4)$. Noting that $\frac{1}{2}(\chi_{[0,1)} + \chi_{[3,4)}) =_{\lambda_{Leb}} \frac{1}{2}\chi_{[0,2)} - \frac{1}{2}\chi_{[1,3)} + \frac{1}{2}\chi_{[2,4)}$ and the set $\{\frac{1}{2}\chi_{[0,2)}, \frac{1}{2}\chi_{[1,3)}, \frac{1}{2}\chi_{[2,4)}\}$ is linearly independent we obtain the independent functional decomposition:

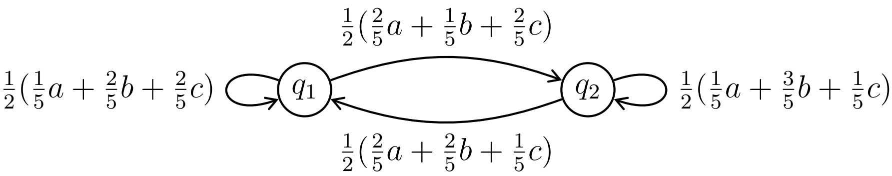
$$\Psi' = \frac{1}{2}\chi_{[0,2)}\begin{pmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix} + \frac{1}{2}\chi_{(1,3)}\begin{pmatrix}0 & \frac{1}{2}\\ 0 & -\frac{1}{2}\end{pmatrix} + \frac{1}{2}\chi_{(2,4)}\begin{pmatrix}0 & 0\\ \frac{1}{2} & \frac{1}{2}\end{pmatrix}.$$

According to Proposition 3.5.5, $P = \begin{pmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{pmatrix}$. Further we compute $\theta = \frac{1}{2}$ and

$d - \theta = \frac{5}{2}$ and

$$\begin{aligned}
M(a) &= \frac{2}{5}\Big[\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} - \frac{1}{2}\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\Big] &= \begin{pmatrix} \frac{1}{10} & \frac{1}{5} \\ \frac{1}{5} & \frac{1}{10} \end{pmatrix} \\
M(b) &= \frac{2}{5}\Big[\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 0 & \frac{1}{2} \\ 0 & -\frac{1}{2} \end{pmatrix}\Big] &= \begin{pmatrix} \frac{1}{5} & \frac{1}{10} \\ \frac{1}{5} & \frac{3}{10} \end{pmatrix} \\
M(c) &= \frac{2}{5}\Big[\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 0 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}\Big] &= \begin{pmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{1}{10} & \frac{1}{10} \end{pmatrix}.
\end{aligned}$$

The HMM from Proposition 3.5.5 is shown below.



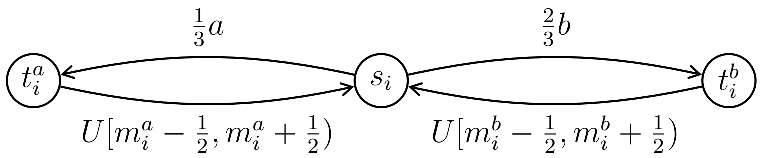
We can compute a basis for $\mathrm{span}\{M(u)\overline{1}^\top \mid u \in \tilde{\Sigma}^*\} = \mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in \mathring{\Sigma}^*\}$ using the algorithm from Proposition 3.1.2 as follows.

Starting with $\mathcal{B} = \{\overline{1}^\top\}$, we first check whether $M(a)\overline{1}^\top \in \mathrm{span}\{\overline{1}^\top\}$. Since $M(a)\overline{1}^\top = (\frac{3}{10}, \frac{3}{10})^\top \in \mathrm{span}\{\overline{1}^\top\}$ we do not add $(\frac{3}{10}, \frac{3}{10})^\top$ to $\mathcal{B}$. We then check whether $M(b)\overline{1}^\top \in \mathrm{span}\{\overline{1}^\top\}$. Since $M(b)\overline{1}^\top = (\frac{3}{10}, \frac{5}{10})^\top \notin \mathrm{span}\{\overline{1}^\top\}$ we set $\mathcal{B} = \{\overline{1}^\top, (\frac{3}{10}, \frac{5}{10})^\top\}$. Since $|\mathcal{B}| = |Q| = 2$ the algorithm terminates.

For any two initial distributions π_1, π_2 may check whether $\pi_1 \equiv \pi_2$ using Theorem 3.3.1. (In this example $\pi_1 \equiv \pi_2$ holds only if $\pi_1 = \pi_2$.) □

3.6 Timing Attacks

Example 3.6.1. We also discuss an example, inspired from [33], where HMM non-equivalence means susceptibility to timing attacks, and HMM equivalence means immunity to such attacks. Consider a system that emits two kinds of observations, both visible to an attacker: a function to be executed (we arbitrarily assume a choice between two functions a and b, and impute a probability distribution between them) and the time it takes to execute that function. An attacker therefore sees a sequence $\ell_1 t_1 \ell_2 t_2 \ldots$, where $\ell_i \in \{a, b\}$ and $t_i \in [0, \infty)$. In [33] the times $t_1, t_2, \ldots$ are all identical and depend only on the secret key held by the system, but we assume in the following that the t_i are drawn from a probability distribution that depends on the function (a or b) and the key. We assume that with key i the execution times have uniform distributions $U[m_i^a - \frac{1}{2}, m_i^a + \frac{1}{2})$ and $U[m_i^b - \frac{1}{2}, m_i^b + \frac{1}{2})$. The situation can then be modelled with the HMM below.[1]

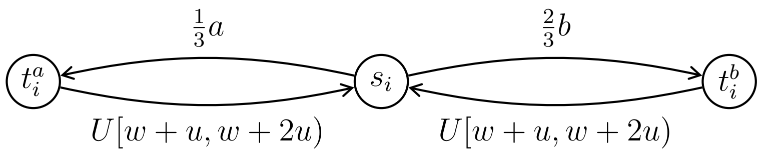


A *timing leak* occurs if the attacker can glean the key from the execution times. For example, the attacker can distinguish between keys k_1 and k_2 if and only if states s_1 and s_2 are not equivalent. One can check, using the algorithm we have developed in this section, that s_1 and s_2 are equivalent if and only if $m_1^a = m_2^a$ and $m_1^b = m_2^b$. Moreover, it follows from section 3.3 that if instead of $U[m_1^a - \frac{1}{2}, m_1^a + \frac{1}{2})$ and $U[m_2^a - \frac{1}{2}, m_2^a + \frac{1}{2})$ we had two distributions with density functions from $[\![\Gamma_{GEM}]\!]$ with the same mean and the same variance, states s_1, s_2 would still be non-equivalent whenever the two distributions are not identical.

One may try to guard against this timing leak by "*padding*" the execution time, so that the sum of the execution time and an added time is constant (and independent of the key). After the execution of the function, an idling loop would be executed until the worst-case (among all keys) execution time of the functions has been reached or exceeded. Let us call this worst-case execution time $w \in (0, \infty)$. This idling loop would take time $u > 0$ in each iteration, so the total idling time is always an integer multiple of u. It is argued in [33] that this guard is in general ineffective in that the attacker can still glean the execution time modulo u. Therefore, it is suggested in [33] to add, in addition, a time that is uniformly distributed on $[0, u)$.

This remedy also works in our case with random execution times. Indeed, one can show that for any independent random variables X, Y, where Y is distributed with $U[0, u]$, we have that $(X + Y) \bmod u$ is distributed with $U[0, u)$. Therefore, by adding an independent $U[0, u)$ random time to the padding described above, the times observable by the attacker now have a $U[w + u, w + 2u)$ distribution, independent of the key.

[1]In this case the observation set $\Sigma = [0, \infty) \cup \{a, b\}$ is a disjoint union of topological spaces and there is a natural measure space induced from the Lebesgue measure space on $[0, \infty)$ and a discrete measure on $\{a, b\}$.



All states s_i are now equivalent, so the key does not leak. $\square$

4
Distinguishability

Definition 4.0.1. Suppose $\mathcal{H} = (Q, \Sigma, \Psi)$ is a HMM and Σ is part of the measure space $(\Sigma, \mathscr{G}, \lambda)$. The *total variation distance* (TV-distance) between two initial distributions π_1 and π_2 is given as

$$d_{\mathcal{H}}(\pi_1, \pi_2) = \sup_{E \in \mathscr{G}^\omega} |\mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)|.$$

Observe that $\pi_1 \equiv_{\mathcal{H}} \pi_2$ if and only if $d_{\mathcal{H}}(\pi_1, \pi_2) = 0$. We sometimes write $d(\pi_1, \pi_2)$ if the choice of HMM is clear. We say π_1 and π_2 are *distinguishable* if $d_{\mathcal{H}}(\pi_1, \pi_2) = 1$.

In this chapter we explore distinguishability in continuous-observation HMMs. In Section 4.1, we give details on the algorithm introduced in [23] for the following decision problem.

Theorem 4.0.1. *Given a finite-observation HMM $\mathcal{C} = (Q, \Sigma, M)$ and initial distributions $\pi_1, \pi_2 \in (\mathbb{Q} \cap [0, 1])^Q$ it is decidable in polynomial time (in the size of the encoding) whether $d(\pi_1, \pi_2) = 1$.*

In Section 4.2 we outline the continuous-observation case. In the finite-observation case the distinguishability decision problem is given as a linear program which is constructed from a basis of the state distribution span (see Proposition 3.1.1) and the *joint-reachability set*. This reachability set also has an analogue in the continuous-observation case. We introduce *joint graph decomposable profile* *languages* which specify sufficient conditions for the joint-reachability set to be efficiently computable: see Section 4.3.

In Section 4.4 we introduce *finite partition approximations* which are finite-observation HMMs that approximate a continuous-observation HMM. For a continuous-observation HMM we show that there exists a finite partition approximation such that any two initial distributions are distinguishable in the continuous-observation HMM if and only if they are distinguishable in the finite partition approximation. While this finite partition approximation is not necessarily computable, it has the same joint-reachability set as the continuous-observation HMM and also the same state distribution span. Thus to construct an analogous linear program from the finite-observation case, we need only compute the joint-reachability set and a basis for the state distribution span. In Proposition 3.3.2, we showed that a basis for the state distribution span can be computed in polynomial time when the profile language is linearly decomposable. Hence, if a HMM is encoded by a linearly decomposable and joint graph decomposable profile language, we may decide in polynomial time whether any two initial distributions are distinguishable.

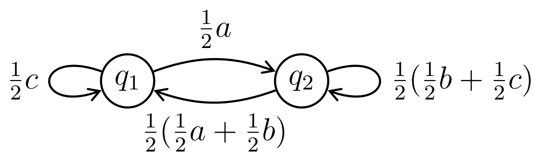
4.1 Distinguishability in the Finite-Observation Case

Let (Q, Σ, M) be a finite-observation HMM and let π_1, π_2 be initial distributions. We define the *joint-reachability set*:

$$R^{\pi_1,\pi_2} = \{(q_1, q_2) \in Q \times Q \mid \exists\, w \in \Sigma^* : (q_1, q_2) \in \operatorname{supp}(\pi_1 M(w)) \times \operatorname{supp}(\pi_2 M(w))\}\,.$$

The joint-reachability set is the joint set of states that are reachable at the same time by π_1 and π_2.

Example 4.1.1. The diagram below shows a finite-observation HMM $\mathcal{F} = (Q, \Sigma, M)$.



Recall that e_{q_1} and e_{q_2} are Dirac distributions on q_1 and q_2 respectively. We have that

$$\begin{aligned}
&\operatorname{supp}(e_{q_1}M(a)) \times \operatorname{supp}(e_{q_2}M(a)) = \{(q_2,q_1)\},\\
&\operatorname{supp}(e_{q_1}M(b)) \times \operatorname{supp}(e_{q_2}M(b)) = \emptyset,\\
&\operatorname{supp}(e_{q_1}M(c)) \times \operatorname{supp}(e_{q_2}M(c)) = \{(q_1,q_2)\},\\
&\operatorname{supp}(e_{q_2}M(a)) \times \operatorname{supp}(e_{q_1}M(a)) = \{(q_1,q_2)\},\\
&\operatorname{supp}(e_{q_2}M(b)) \times \operatorname{supp}(e_{q_1}M(b)) = \emptyset,\\
&\operatorname{supp}(e_{q_2}M(c)) \times \operatorname{supp}(e_{q_1}M(c)) = \{(q_2,q_1)\}.
\end{aligned} \tag{4.1}$$

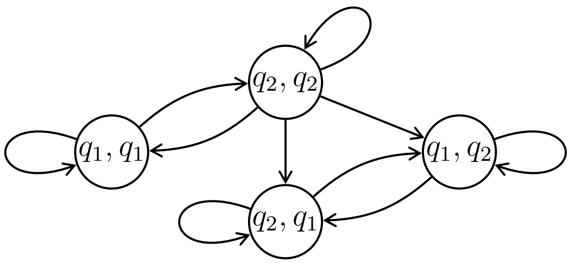
It follows by the first and fifth equalities that $\operatorname{supp}(e_{q_1}M(ac)) \times \operatorname{supp}(e_{q_2}M(ac)) = \{(q_2,q_1)\}$. Further, for any $w \in \Sigma^*$ we have $\operatorname{supp}(e_{q_1}M(w)) \times \operatorname{supp}(e_{q_2}M(w))$ contains at most one pair of states: either (q_1,q_2) or (q_2,q_1). Thus, $R^{e_{q_1},e_{q_2}} = \{(q_1,q_2),(q_2,q_1)\}$.

We can rephrase the equations in (4.1) as transitions between pairs of states. We define the *joint graph*:

Definition 4.1.1. Let $\mathcal{F} = (Q,\Sigma,M)$ be a finite-observation HMM. The joint graph is $G_{\mathcal{F},\mathcal{F}} = (Q \times Q, E)$ where

$$E = \{((q_1,r_1),(q_2,r_2)) \mid \exists\, a \in \Sigma : M(a)_{q_1,q_2} > 0 \text{ and } M(a)_{r_1,r_2} > 0\}.$$

Example 4.1.2. Continuing with the HMM $\mathcal{F}$ from Example 4.1.1, the diagram below shows the joint graph $G_{\mathcal{F},\mathcal{F}}$.



We see that the first equation in (4.1) implies that there is an edge from (q_1,q_2) to (q_2,q_1). Additionally we have $\operatorname{supp}(e_{q_2}M(b)) = Q$ and so $\operatorname{supp}(e_{q_2}M(b)) \times \operatorname{supp}(e_{q_2}M(b)) = Q \times Q$. Indeed, we see in the graph that from (q_2,q_2) there is an edge to every vertex in $G_{\mathcal{F},\mathcal{F}}$.

There is a link between R^{π_1,π_2} and the reachable states from $\operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2)$ in $G_{\mathcal{F},\mathcal{F}}$. Recall from the preliminaries that for $U, V \subseteq Q \times Q$ we write $U \to_{G_{\mathcal{F},\mathcal{F}}} V$ if there exists $u \in U$ and $v \in V$ such that v is reachable from u in $G_{\mathcal{F},\mathcal{F}}$. We have the following which is Lemma 20 from [23].

Lemma 4.1.3. *Let $\mathcal{F} = (Q, \Sigma, M)$ be a finite-observation HMM and let π_1, π_2 be initial distributions. We have*

$$R^{\pi_1,\pi_2} = \left\{(r_1, r_2) \in Q \times Q \;\middle|\; \mathit{supp}\,(\pi_1) \times \mathit{supp}\,(\pi_2) \to_{G_{\mathcal{F},\mathcal{F}}} \{(r_1, r_2)\}\right\}.$$

As a consequence R^{π_1,π_2} can be computed in polynomial time using graph reachability.

Example 4.1.4. We again continue with $\mathcal{F}$ from Example 4.1.1 where we computed $R^{e_{q_1},e_{q_2}}$. Using the graph calculated in Example 4.1.2 we have that the set of reachable vertices from (q_1, q_2) is $\{(q_1, q_2), (q_2, q_1)\}$ which indeed coincides with $R^{e_{q_1},e_{q_2}}$.

On the other hand, every vertex in $Q \times Q$ is reachable from (q_1, q_1) and so $R^{e_{q_1},e_{q_1}} = Q \times Q$.

We now state the main result used in deciding whether two initial distributions are distinguishable. This result is Theorem 21 from [23].

Theorem 4.1.5. *Let (Q, Σ, M) be a finite-observation HMM and let π_1, π_2 be initial distributions. For each $r_1 \in Q$ we define the projection*

$$R^{\pi_1,\pi_2}_{r_1} := \{r_2 \in Q \mid (r_1, r_2) \in R^{\pi_1,\pi_2}\}.$$

Then $d(\pi_1, \pi_2) < 1$ holds if and only if there is $r_1 \in Q$ and subdistributions μ_1, μ_2 such that

$$\mu_1 \equiv \mu_2 \text{ and } r_1 \in \mathit{supp}\,(\mu_1) \text{ and } \mathit{supp}\,(\mu_2) \subseteq R^{\pi_1,\pi_2}_{r_1}. \tag{4.2}$$

We may check the condition $\mu_1 \equiv \mu_2$ using the basis computed in the following algorithm from Section 3.1.

Proposition 3.1.2. *Let (Q, Σ, Ψ) be a finite-observation HMM. One can compute a basis $\mathcal{B}$ for* $\operatorname{span}\{\Psi(w)\bar{1}^\top \mid w \in \Sigma^*\}$ *in polynomial time.*

Example 4.1.6. Taking $\mathcal{F}$ from Example 4.1.1 we have

$$M(a) = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{4} & 0 \end{pmatrix}, \quad M(b) = \begin{pmatrix} 0 & 0 \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} \text{ and } M(c) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}.$$

We have $M(a)\overline{1}^\top = (\frac{1}{2}, \frac{1}{4})^\top$ and $M(b)\overline{1}^\top = (0, \frac{1}{2})^\top$. The set $\{(\frac{1}{2}, \frac{1}{4})^\top, (0, \frac{1}{2})^\top\}$ is linearly independent and so $\operatorname{span}\{M(w)\overline{1}^\top \mid w \in \Sigma^*\} = \mathbb{R}^2$. It follows that for two subdistributions μ_1, μ_2 we have $\mu_1 - \mu_2 \perp \operatorname{span}\{M(w)\overline{1}^\top \mid w \in \Sigma^*\}$ if and only if $\mu_1 = \mu_2$.

Consider $R_{q_1}^{e_{q_1}, e_{q_2}} = \{q_2\}$. There is no pair of subdistributions μ_1, μ_2 satisfying (4.2) since $\mu_1 = \mu_2$ implies that

$$q_1 \in \operatorname{supp}(\mu_1) \text{ and } \operatorname{supp}(\mu_2) \subseteq R_{q_1}^{e_{q_1}, e_{q_2}} = \{q_2\}$$

contradict each other. A similar contradiction arises when considering $R_{q_2}^{e_{q_1}, e_{q_2}} = \{q_1\}$. It follows by Theorem 4.1.5 that $d(e_{q_1}, e_{q_2}) = 1$.

With the basis $\mathcal{B}$ from Proposition 3.1.2, the condition $\mu_1 \equiv \mu_2$ is equivalent to $\forall b \in \mathcal{B} \quad (\mu_1 - \mu_2) \cdot b = 0$. We can decide in polynomial time whether there exists $\mu_1, \mu_2 \in \mathbb{Q}^{|Q|}$ such that $q \in \operatorname{supp}(\mu_1)$ and $\operatorname{supp}(\mu_2) \subseteq R_{r_1}^{\pi_1, \pi_2}$ and $\forall b \in \mathcal{B} \quad (\mu_1 - \mu_2) \cdot b = 0$ using linear programming. The following algorithm is inspired by Algorithm 1 from [23].

Algorithm 1 An algorithm for deciding distance 1 in the finite-observation case

compute R^{π_1, π_2} by graph reachability (Lemma 4.1.3)
compute $\mathcal{B} \subseteq \mathbb{Q}^{|Q|}$ from Proposition 3.1.2
for $r_1 \in Q$ **do**
  if $\exists\ \mu_1, \mu_2 \in \mathbb{Q}^{|Q|}$ s.t. $r_1 \in \operatorname{supp}(\mu_1)$ and $\operatorname{supp}(\mu_2) \subseteq R_{r_1}^{\pi_1, \pi_2}$ and
    $\forall b \in \mathcal{B} \quad (\mu_1 - \mu_2) \cdot b = 0$ **then**
    return "$d(\pi_1, \pi_2) < 1$"
  end if
end for
return "$d(\pi_1, \pi_2) = 1$"

Theorem 4.0.1 follows immediately from Algorithm 1. However, in [23] the algorithm is slightly different. They compute a basis $\mathcal{B}'$ for

$$V := \operatorname{span}\left\{ \begin{pmatrix} M(w) & 0 \\ 0 & -M(w) \end{pmatrix} \overline{1}^\top \,\middle|\, w \in \Sigma^* \right\}.$$

They then check the condition $(\mu_1 \ \mu_2) \cdot b' = 0$ holds for all $b' \in \mathcal{B}'$. Here, $(\mu_1 \ \mu_2) \in [0,1]^{2|Q|}$ is the row vector obtained by gluing μ_1, μ_2 together. Note that $(\mu_1 \ \mu_2) \cdot b$ is a scalar.

The condition $(\mu_1 \ \mu_2) \cdot b' = 0$ holds for all $b' \in \mathcal{B}'$ is equivalent to $(\mu_1 - \mu_2) \cdot b = 0$ for all $b \in \mathcal{B}$ (See Lemma A.2.1 in the appendix). We use the latter condition because later in this chapter, when we generalise to continuous-observation HMMs we will reuse results presented in Chapter 3 in terms of $\mathrm{span}\{M(w)\overline{1}^\top \mid w \in \Sigma^*\}$.

We now compare the finite-observation case with the continuous-observation case.

4.2 Distinguishability in the Continuous-Observation Case

In the following we present an algorithm to decide whether two initial distributions are distinguishable when the HMM has continuous observations. In particular there is an analogue of Theorem 4.1.5 which leads to an analogue of Algorithm 1. First we define the *continuous joint-reachability set*.

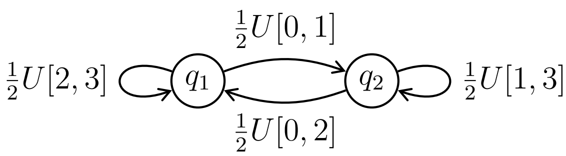
Definition 4.2.1. Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a HMM and let π_1, π_2 be initial distributions.

$$\tilde{R}^{\pi_1,\pi_2} = \Big\{(q_1, q_2) \in Q \times Q \mid \exists n \in \mathbb{N}_0 : \\ \lambda^n\Big(w \in \Sigma^n \mid (q_1, q_2) \in \mathrm{supp}\,(\pi_1\Psi(w)) \times \mathrm{supp}\,(\pi_2\Psi(w))\Big) > 0\Big\}.$$

Here, λ^n is the measure on Σ^n as defined in the preliminaries. We write λ^0 for the measure on $\Sigma^0 = \{\varepsilon\}$ where $\lambda^0(\{\varepsilon\}) = 1$. It follows that

$$\mathrm{supp}\,(\pi_1) \times \mathrm{supp}\,(\pi_2) \subseteq \tilde{R}^{\pi_1,\pi_2}. \tag{4.3}$$

Example 4.2.1. Consider a continuous variation of Example 4.1.1 shown below.



In the HMM above $U[a,b]$ is a uniform distribution on the closed interval $[a,b]$ and $\lambda = \lambda_{Leb}$: the Lebesgue measure on $\mathbb{R}$. We compute $\tilde{R}^{e_{q_1},e_{q_2}}$.

By (4.3) we have $\operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2) = (q_1, q_2) \subseteq \tilde{R}^{\pi_1,\pi_2}$. Also, $(q_2, q_1) \in \operatorname{supp}(e_{q_1}\Psi(x)) \times \operatorname{supp}(e_{q_2}\Psi(x))$ for all $x \in [0,1)$ and $\lambda_{Leb}([0,1)) = 1 > 0$ which implies $(q_2, q_1) \subseteq \tilde{R}^{\pi_1,\pi_2}$.

Further we have that

$$\operatorname{supp}(e_{q_1}\Psi(x)) \times \operatorname{supp}(e_{q_2}\Psi(x)) = \begin{cases} \{(q_2,q_1)\} & x \in [0,1) \\ \{(q_2,q_1),(q_2,q_2)\} & x = 1 \\ \emptyset & x \in (1,2) \\ \{(q_1,q_1),(q_1,q_2)\} & x = 2 \\ \{(q_1,q_2)\} & x \in (2,3] \end{cases}$$

and

$$\operatorname{supp}(e_{q_2}\Psi(x)) \times \operatorname{supp}(e_{q_1}\Psi(x)) = \begin{cases} \{(q_1,q_2)\} & x \in [0,1) \\ \{(q_2,q_1),(q_2,q_2)\} & x = 1 \\ \emptyset & x \in (1,2) \\ \{(q_1,q_1),(q_2,q_1)\} & x = 2 \\ \{(q_2,q_1)\} & x \in (2,3]. \end{cases}$$

and so any set of words which leads to states (q_1, q_1) or (q_2, q_2) is λ_{Leb}-null. Therefore $\tilde{R}^{\pi_1,\pi_2} = \{(q_1,q_2),(q_2,q_1)\}$.

There is an analogue of the joint graph for continuous observation HMMs. Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a continuous-observation HMM. Recall that for a function $f : \Sigma \to [0,\infty)$, $\operatorname{supp}(f) = \{x \in \Sigma \mid f(x) \neq 0\}$. We write $\tilde{G}_{\mathcal{C},\mathcal{C}} = (Q \times Q, \tilde{E})$ where

$$\tilde{E} = \{((q_1,r_1),(q_2,r_2)) \in (Q \times Q) \times (Q \times Q) \mid \lambda(\operatorname{supp}(\Psi_{q_1,q_2}) \cap \operatorname{supp}(\Psi_{r_1,r_2})) > 0\}.$$

Example 4.2.2. We continue with the HMM $\mathcal{C}$ from Example 4.2.1 and look at some examples of edges in $\tilde{G}_{\mathcal{C},\mathcal{C}}$. We have

$$\lambda(\operatorname{supp}(\Psi_{q_2,q_1}) \cap \operatorname{supp}(\Psi_{q_2,q_2})) = \lambda([0,2] \cap [1,3]) = \lambda([0,1]) = 1$$

and so there is an edge from (q_2, q_2) to (q_1, q_2). We have

$$\lambda(\operatorname{supp}(\Psi_{q_1,q_1} \cap \operatorname{supp}\Psi_{q_2,q_1})) = \lambda([2,3] \cap [0,2]) = \lambda(\{0\}) = 0$$

and so there is no edge from (q_1, q_2) to (q_1, q_1). Continuing in this fashion we can show that $\tilde{G}_{\mathcal{C},\mathcal{C}} = G_{\mathcal{F},\mathcal{F}}$ where $\mathcal{F}$ is the finite-observation HMM from Example 4.1.1.

Towards an analogue of Lemma 4.1.3 we would like to compute the joint graph of a HMM with continuous observations. Recall that a continuous-observation HMM is over a specific profile language and is encoded by list of coefficient-profile pairs. Each profile specifies the observation distribution of a particular transition in the HMM. Computing the edge set of the joint graph requires deciding whether any two profiles encode functions with a shared non-null region of support. We specify the computability of this decision problem as a property of the profile language. We call such a profile language *joint graph decomposable.*

Definition 4.2.2. Let Γ be a profile language with $[\![\Gamma]\!] \subseteq \mathcal{L}_1(\Sigma, \lambda)$. We say that Γ is joint graph decomposable if there is a procedure that takes a pair of profiles $\alpha, \beta \in \Gamma$ and decides whether $\lambda(\mathrm{supp}\,([\![\alpha]\!]) \cap \mathrm{supp}\,([\![\beta]\!])) > 0$ and this procedure takes polynomial time in the combined length of the encodings of α and β.

Example 4.2.3. Let Γ be a profile language encoding functions in the set

$$\{f : \mathbb{R} \to [0, \infty) \mid \exists\, a, b \in \mathbb{Q} \cup \{-\infty, \infty\} : \mathrm{supp}\,(f) = (a, b) \text{ up to a } \lambda_{Leb}\text{-null set}\}. \tag{4.4}$$

For any $\gamma \in \Gamma$ we have $\mathrm{supp}\,[\![\gamma]\!] = (a, b)$ for some $a, b \in \mathbb{Q} \cup \{-\infty, \infty\}$. Assume a and b are included in γ as part of the encoding. Then, let $\gamma_1, \gamma_2 \in \Gamma$ such that $\mathrm{supp}\,([\![\gamma_i]\!]) = (a_i, b_i)$ for $i = 1, 2$. Then we have $\lambda(\mathrm{supp}\,([\![\gamma_1]\!]) \cap \mathrm{supp}\,([\![\gamma_2]\!])) > 0$ if and only if both $a_1 < b_2$ and $a_2 < b_1$. The latter is decidable in constant time. The set of functions (4.4) includes most standard probability density functions. We show later in Chapter 5 that a large class of standard probability density functions is encodable by a joint graph decomposable profile language.

Recall the profile language Γ_{GEM} from Section 3.3. We prove the following later in Section 5.3.

Proposition 4.2.4. *The profile language Γ_{GEM} is joint graph decomposable.*

Additionally, the next result shows that joint graph decomposability is a sufficient condition for being able to compute the joint graph.

Proposition 4.2.5. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a HMM over a joint graph decomposable language then $\tilde{G}_{\mathcal{C},\mathcal{C}}$ can be computed in polynomial time.*

Proof. Suppose the HMM $\mathcal{C}$ is over the joint graph decomposable profile language Γ. From the description of Ψ as a matrix of coefficient-profile pairs, we have $\Psi_{i,j} = p_{i,j}[\![\gamma_{i,j}]\!]$ for all $i, j \in [Q]$ where $p_{i,j} \in \mathbb{Q}^+$ and $\gamma_{i,j} \in \Gamma$.

To compute $\tilde{G}_{\mathcal{C},\mathcal{C}}$, we need only compute its set of edges. For each $i_1, i_2, j_1, j_2 \in [Q]$ the decision problem

$$\lambda(\mathrm{supp}\,([\![\gamma_{i_1,i_2}]\!]) \cap \mathrm{supp}\,([\![\gamma_{j_1,j_2}]\!])) > 0$$

is decidable in polynomial time because Γ is joint graph decomposable. It follows that the construction of $\tilde{G}_{\mathcal{C},\mathcal{C}}$ takes polynomial time. $\square$

As a consequence of the above result we have an analogue of Lemma 4.1.3 and subsequently an analogue of Theorem 4.1.5.

Lemma 4.2.6. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a continuous-observation HMM over a joint graph decomposable profile language. Let π_1, π_2 be initial distributions. We have*

$$\tilde{R}^{\pi_1,\pi_2} = \left\{(r_1, r_2) \in Q \times Q \;\middle|\; \mathit{supp}\,(\pi_1) \times \mathit{supp}\,(\pi_2) \rightarrow_{\tilde{G}_{\mathcal{C},\mathcal{C}}} \{(r_1, r_2)\}\right\}.$$

As a consequence $\tilde{R}^{\pi_1,\pi_2}$ can be computed in polynomial time using graph reachability.

Theorem 4.2.7. *Let (Q, Σ, Ψ) be a continuous-observation HMM and let π_1, π_2 be initial distributions. For each $r_1 \in Q$ we define the projection*

$$\tilde{R}^{\pi_1,\pi_2}_{r_1} := \{r_2 \in Q \mid (r_1, r_2) \in \tilde{R}^{\pi_1,\pi_2}\}.$$

Then $d(\pi_1, \pi_2) < 1$ holds if and only if there are $r_1 \in Q$ and subdistributions μ_1, μ_2 such that

$$\mu_1 \equiv \mu_2 \textit{ and } r_1 \in \mathit{supp}\,(\mu_1) \textit{ and } \mathit{supp}\,(\mu_2) \subseteq \tilde{R}^{\pi_1,\pi_2}_{r_1}. \tag{4.5}$$

In a similar vein to the finite-observation case, we can replace the equivalence condition in (4.5) with orthogonality to a particular basis.

Corollary 4.2.8. *Let (Q, Σ, Ψ) be a continuous-observation HMM and let π_1, π_2 be initial distributions. Suppose Ψ has set of continuity $\mathring{\Sigma}$. Then, $d(\pi_1, \pi_2) < 1$ holds if and only if there are $r_1 \in Q$ and subdistributions μ_1, μ_2 such that*

$$\forall\, b \in \tilde{\mathcal{B}} \quad (\mu_1 - \mu_2) \cdot b = 0 \textit{ and } r_1 \in \mathit{supp}\,(\mu_1) \textit{ and } \mathit{supp}\,(\mu_2) \subseteq \tilde{R}^{\pi_1,\pi_2}_{r_1}$$

where $\tilde{\mathcal{B}}$ is a basis for $\mathrm{span}\{\Psi(w)\overline{1}^\top \mid w \in \mathring{\Sigma}^*\}$.

Proof. By Proposition 3.1.5 two subdistributions μ_2, μ_2 are equivalent if and only if $(\mu_1 - \mu_2) \perp \operatorname{span}\{\Psi(w)\bar{1}^\top \mid w \in \mathring{\Sigma}^*\}$ which holds if and only if $(\mu_1 - \mu_2) \cdot b = 0$ for all $b \in \tilde{\mathcal{B}}$. □

Recall from Chapter 3 that the basis $\tilde{\mathcal{B}}$ from the previous corollary can be computed in polynomial time:

Proposition 3.3.2. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a continuous-observation HMM over a linearly decomposable profile language. Suppose Ψ has set of continuity $\mathring{\Sigma}$. We may compute a basis $\mathcal{B}$ for* $\operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\}$ *in polynomial time (in the size of the encoding of $\mathcal{C}$).*

We have now assembled the ingredients of a polynomial time algorithm to decide whether two initial distributions are distinguishable in a continuous-observation HMM.

Algorithm 2 An algorithm for deciding distance 1 in the continuous-observation case

```
compute R̃^{π1,π2} by graph reachability (Lemma 4.2.6)
compute B̃ ⊆ Q^{|Q|} from Proposition 3.3.2
for r1 ∈ Q do
    if ∃ μ1, μ2 ∈ Q^{|Q|} s.t. r1 ∈ supp μ1 and supp μ2 ⊆ R̃_{r1}^{π1,π2} and
            ∀b ∈ B̃   (μ1 − μ2) · b = 0 then
        return "d(π1, π2) < 1"
    end if
end for
return "d(π1, π2) = 1"
```

Theorem 4.2.9. *Given a HMM $\mathcal{C} = (Q, \Sigma, \Psi)$ over a linearly decomposable and joint graph decomposable profile language and initial distributions $\pi_1, \pi_2 \in (\mathbb{Q} \cap [0,1])^Q$ it is decidable in polynomial time (in the size of the encoding) whether $d(\pi_1, \pi_2) = 1$.*

Proof. The basis $\tilde{\mathcal{B}}$ from Proposition 3.3.2 can be computed in polynomial time. By Lemma 4.2.6 the continuous joint-reachability set $\tilde{R}^{\pi_1,\pi_2}$ can be computed in polynomial time. By Corollary 4.2.8 it suffices to check

$$\begin{aligned} \exists\, \mu_1, \mu_2 \text{ s.t. } \forall\, b \in \tilde{\mathcal{B}} \quad & (\mu_1 - \mu_2) \cdot b = 0 \text{ and} \\ & r_1 \in \operatorname{supp}(\mu_1) \text{ and} \\ & \operatorname{supp}(\mu_2) \subseteq \tilde{R}_{r_1}^{\pi_1,\pi_2}. \end{aligned}$$

which is decidable in polynomial time by linear programming. □

By Proposition 4.2.4 we have:

Corollary 4.2.10. *Given a HMM $\mathcal{H} = (Q, \Sigma, \Psi)$ over Γ_{GEM}, and initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^{|Q|}$, it is decidable in polynomial time whether $d_{\mathcal{H}}(\pi_1, \pi_2) = 1$.*

The next sections are dedicated to the missing proofs of the results in this section.

4.3 The Joint-Reachability Set

We first restate the following definition.

Definition 4.2.1. Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a HMM and let π_1, π_2 be initial distributions.

$$\tilde{R}^{\pi_1,\pi_2} = \Big\{(q_1, q_2) \in Q \times Q \mid \exists n \in \mathbb{N}_0 : \\ \lambda^n\Big(w \in \Sigma^n \mid (q_1, q_2) \in \operatorname{supp}(\pi_1 \Psi(w)) \times \operatorname{supp}(\pi_2 \Psi(w))\Big) > 0\Big\}.$$

Then, taking an alternative approach to the proof of Lemma 4.1.3 found in [23] we prove the following.

Lemma 4.2.6. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a continuous-observation HMM over a joint graph decomposable profile language. Let π_1, π_2 be initial distributions. We have*

$$\tilde{R}^{\pi_1,\pi_2} = \Big\{(r_1, r_2) \in Q \times Q \;\Big|\; \mathit{supp}\,(\pi_1) \times \mathit{supp}\,(\pi_2) \to_{\tilde{G}_{\mathcal{C},\mathcal{C}}} \{(r_1, r_2)\}\Big\}.$$

As a consequence $\tilde{R}^{\pi_1,\pi_2}$ can be computed in polynomial time using graph reachability.

Proof. Let $n \in \mathbb{N}$ and $q_0, \ldots, q_n \in Q$ and $r_0, \ldots, r_n \in Q$ then

$$\begin{aligned}
&\lambda^n\Big(\{a_1 \cdots a_n \in \Sigma^n \mid \Psi(a_1)_{q_0,q_1} \cdots \Psi(a_1)_{q_{n-1},q_n} > 0 \text{ and} \\
&\qquad\qquad \Psi(a_1)_{r_0,r_1} \cdots \Psi(a_1)_{r_{n-1},r_n} > 0\Big) > 0 \\
\iff &\lambda^n\Big(\{a_1 \cdots a_n \in \Sigma^n \mid \Psi(a_i)_{q_{i-1},q_i} > 0 \text{ and} \\
&\qquad\qquad \Psi(a_i)_{r_{j-1},r_j} > 0 \quad \forall i \in [n]\}\Big) > 0 \\
\iff &\prod_{i=1}^{n} \lambda(a \in \Sigma \mid \Psi(a)_{q_{i-1},q_i} > 0 \text{ and } \Psi(a)_{r_{i-1},r_i} > 0) > 0 \\
\iff &\prod_{i=1}^{n} \lambda(\operatorname{supp}\Psi_{q_{i-1},q_i} \cap \operatorname{supp}\Psi_{r_{i-1},r_i}) > 0 \\
\iff &\lambda(\operatorname{supp}\Psi_{q_{i-1},q_i} \cap \operatorname{supp}\Psi_{r_{i-1},r_i}) > 0 \quad \forall\, i \in [n] \\
\iff &(q_0, r_0) \cdots (q_n, r_n) \text{ is a path in } \tilde{G}_{\mathcal{C},\mathcal{C}}.
\end{aligned} \tag{4.6}$$

Define

$$S^{\pi_1,\pi_2} = \left\{ (r_1, r_2) \in Q \times Q \;\middle|\; \operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2) \to_{\tilde{G}_{\mathcal{C},\mathcal{C}}} \{(r_1, r_2)\} \right\}.$$

Let $(q, r) \in \tilde{R}^{\pi_1,\pi_2}$ then there exists $n \in \mathbb{N}_0$ such that

$$\lambda^n \Big(w \in \Sigma^n \mid (q_1, q_2) \in \operatorname{supp}(\pi_1 \Psi(w)) \times \operatorname{supp}(\pi_2 \Psi(w)) \Big) > 0.$$

In the case $n = 0$, we have $(q, r) \in \operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2)$ which is clearly reachable from itself. Hence, $(q, r) \in S^{\pi_1,\pi_2}$.

In the case $n \in \mathbb{N}$ then there is $q_0, \ldots, q_n \in Q$ and $r_0, \ldots, r_n \in Q$ such that $q_0 \in \operatorname{supp} \pi_1$, $r_0 \in \operatorname{supp} \pi_2$, $q_n = q$, $r_n = r$ and

$$\lambda^n \Big(\{ a_1 \cdots a_n \in \Sigma^n \mid \Psi(a_i)_{q_{i-1},q_i} > 0 \text{ and } \Psi(a_i)_{r_{j-1},r_j} > 0 \quad \forall i \in [n] \} \Big) > 0.$$

By (4.6) it follows that there is a path $(q_0, r_0), \ldots, (q_n, r_n)$ in the graph $\tilde{G}_{\mathcal{C},\mathcal{C}}$. Thus, $(q, r) \in S^{\pi_1,\pi_2}$.

Now assume $(q, r) \in S^{\pi_1,\pi_2}$. There is a path $(q_0, r_0), \ldots, (q_n, r_n) \in Q \times Q$ such that $(q_0, r_0) \in \operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2)$ and $(q_n, r_n) = (q, r)$. In the case $n = 0$ then $(q, r) \in \operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2) \subseteq \tilde{R}^{\pi_1,\pi_2}$. In the case $n \in \mathbb{N}$, by (4.6), we have

$$\begin{aligned} 0 < \lambda^n \Big(\{ a_1 \cdots a_n \in \Sigma^n \mid \Psi(a_1)_{q_0,q_1} \cdots \Psi(a_1)_{q_{n-1},q_n} > 0 \text{ and } \\ \Psi(a_1)_{r_0,r_1} \cdots \Psi(a_1)_{r_{n-1},r_n} > 0 \Big) \\ \leq \lambda^n \Big(\{ w \in \Sigma^n \mid (\pi_1 \Psi(w))_{q_n} > 0 \text{ and } (\pi_2 \Psi(w))_{r_n} > 0 \} \Big) \end{aligned}$$

Hence $(q, r) \in \tilde{R}^{\pi_1,\pi_2}$. It follows that $\tilde{R}^{\pi_1,\pi_2} = S^{\pi_1,\pi_2}$. □

4.4 Finite Partition Approximations

This section is dedicated to the proof of Theorem 4.2.7. We begin by defining *$\mathcal{V}$-approximations*:

Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a HMM and suppose Ψ has set of continuity $\mathring{\Sigma}$. Let $\mathcal{V}$ be a finite partition of $\mathring{\Sigma}$. Let $\mathcal{V}' = \{V_1, \ldots, V_K\} := \{V \in \mathcal{V} \mid \int_V \Psi d\lambda \neq 0\}$.

Define a set $\tilde{\Sigma} = \{a_1, \ldots, a_K\}$ of fresh observations and an observation density matrix M with $M(a_k) = \int_{V_k} \Psi d\lambda$. We define the $\mathcal{V}$-approximation of $\mathcal{C}$ as $\mathcal{F} = (Q, \tilde{\Sigma}, M)$.

Lemma 4.4.1. *Let $\mathcal{C}$ be a HMM and let $\mathcal{F}$ be a $\mathcal{V}$-approximation of $\mathcal{C}$. We have that $\mathcal{F}$ is a valid finite-observation HMM and for any initial distributions π_1, π_2*

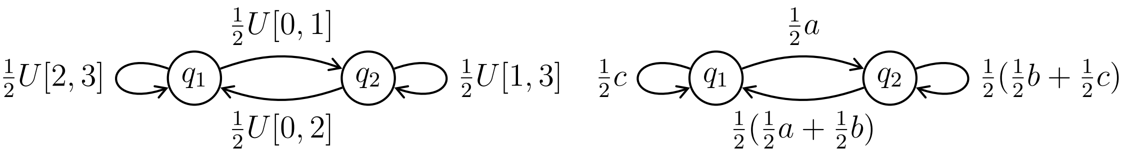
$$d_{\mathcal{C}}(\pi_1, \pi_2) \geq d_{\mathcal{F}}(\pi_1, \pi_2).$$

Before proving the above lemma, we illustrate $\mathcal{V}$-approximations with an example.

Example 4.4.2. Take the HMM from Example 4.2.1. We have $\Sigma = \mathbb{R}$ and $\mathring{\Sigma} = (-\infty, 0) \cup (0, 1) \cup (1, 2) \cup (2, 3) \cup (3, \infty)$. Take $\mathcal{V} = \{(-\infty, 0), (0, 1), (1, 2), (2, 3), (3, \infty)\}$. We have $\mathcal{V}' = \{(0, 1), (1, 2), (2, 3)\}$. Let $\tilde{\Sigma} = \{a, b, c\}$. We define the observation density matrix:

$$\begin{aligned}
M(a) &= \int_0^1 \Psi d\lambda_{Leb} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{4} & 0 \end{pmatrix} \\
M(b) &= \int_1^2 \Psi d\lambda_{Leb} = \begin{pmatrix} 0 & 0 \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} \\
M(c) &= \int_2^3 \Psi d\lambda_{Leb} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}
\end{aligned}$$

The $\mathcal{V}$-approximation $\mathcal{F} = (Q, \tilde{\Sigma}, M)$ is the same as the HMM from Example 4.1.1. We show the HMM from Example 4.2.1 and its $\mathcal{V}$-approximation below.



Proof of Lemma 4.4.1. Since $\Psi(x) \geq 0$ for all $x \in \Sigma$ it follows that for each $V \in \mathcal{V}$ the integral $\int_V \Psi d\lambda$ is non-negative. Further, since $\mathcal{V}$ is a partition, $\sum_{V \in \mathcal{V}} \int_V \Psi d\lambda = \int_\Sigma \Psi d\lambda$ and hence $\sum_{k=1}^K M(a_k)$ is stochastic. Thus, $\mathcal{F}$ defines a HMM.

The σ-algebra $\mathcal{G}_{\mathcal{V}} = \sigma\{V_1, \ldots, V_K\}$ is a subset of $\mathcal{G}$. Let $\mathcal{G}^*_{\mathcal{V}}$ be the induced σ-algebra on infinite words $\tilde{\Sigma}^\omega$, then $\mathcal{G}^*_{\mathcal{V}} \subseteq \mathcal{G}^*$. It follows that

$$d_{\mathcal{F}}(\pi_1, \pi_2) = \sup_{E \in \mathscr{G}^*_{\mathcal{V}}} |\mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)| \leq \sup_{E \in \mathscr{G}^\omega} |\mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)| = d_{\mathcal{H}}(\pi_1, \pi_2). \quad \square$$

Under certain conditions on a $\mathcal{V}$-approximation, the inequality in Lemma 4.4.1 is an equality:

Proposition 4.4.3. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ be a continuous observation HMM. Assume Ψ has set of continuity $\mathring{\Sigma}$. Let $\mathcal{V}$ be a finite partition of $\mathring{\Sigma}$ such that:*

1. $\operatorname{span}\left\{ \int_V \Psi d\lambda \mid V \in \mathcal{V} \right\} = \operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\}$,
2. *For each $V \in \mathcal{V}$ the image $supp\,(\Psi(V))$ contains exactly one element.*

Let $\mathcal{F} = (Q, \tilde{\Sigma}, M)$ be the $\mathcal{V}$-approximation of $\mathcal{C}$ then for any initial distributions π_1, π_2 we have

$$d_{\mathcal{F}}(\pi_1, \pi_2) = 1 \iff d_{\mathcal{H}}(\pi_1, \pi_2) = 1.$$

Towards a proof of Proposition 4.4.3 we state the following lemma which is Proposition 17 from [23].

Lemma 4.4.4. *Let $\mathcal{F} = (Q, \Sigma, M)$ be a finite-observation HMM and let π_1, π_2 be initial distributions. Then $d(\pi_1, \pi_2) < 1$ if and only if there are $w \in \Sigma^*$ and subdistributions μ_1, μ_2 with $\mu_1 \le \pi_1 M(w)$ and $\mu_2 \le \pi_2 M(w)$ and $\mu_1 \equiv_{\mathcal{F}} \mu_2$ and $\|\mu_1\| = \|\mu_2\| > 0$.*

The intuition for Proposition 4.4.3 is that the first point implies that $\pi_1 \equiv_{\mathcal{F}} \pi_2$ if and only if $\pi_1 \equiv_{\mathcal{C}} \pi_2$. Together with the first point, the second point of Proposition 4.4.3 implies that if $d_{\mathcal{F}}(\pi_1, \pi_2) < 1$ then the word w in the statement of Lemma 4.4.4 has a corresponding cylinder set $\overline{E}$ such that there are subdistributions $\overline{\mu_1}, \overline{\mu_2}$ where $\overline{\mu_1} \le \pi_1 \int_{\overline{E}} \Psi d\lambda^{|w|}$ and $\mu_2 \le \pi_2 \int_{\overline{E}} \Psi d\lambda^{|w|}$ and $\overline{\mu_1} \equiv_{\mathcal{C}} \overline{\mu_2}$ and $\|\overline{\mu_1}\| = \|\overline{\mu_2}\| > 0$. We show that this condition implies $d_{\mathcal{C}}(\pi_1, \pi_2) < 1$. Finally, by Lemma 4.4.1, we have $d_{\mathcal{F}}(\pi_1, \pi_2) = 1$ implies $d_{\mathcal{C}}(\pi_1, \pi_2) = 1$.

Example 4.4.5. We show that the partition $\mathcal{V}$ and HMM $\mathcal{C}$ from Example 4.4.2 satisfies the properties in Proposition 4.4.3. We have

$$\Psi(x) = \begin{cases} \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{4} & 0 \end{pmatrix} & x \in (0,1) \\[2em] \begin{pmatrix} 0 & 0 \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} & x \in (1,2) \\[2em] \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} & x \in (2,3). \end{cases}$$

Thus,

$$\begin{aligned}
\operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\} &= \operatorname{span}\left\{\begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{4} & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}\right\} \\
&= \operatorname{span}\left\{M(a), M(b), M(c)\right\} \\
&= \operatorname{span}\left\{\int_V \Psi d\lambda \mid V \in \mathcal{V}\right\}
\end{aligned}$$

which shows that $\mathcal{V}$ satisfies the first condition in Proposition 4.4.3. For the second condition we have the following cases:

$$\operatorname{supp}(\Psi(x)) = \begin{cases} \{(q_1, q_2), (q_2, q_1)\} & x \in (0,1) \\ \{(q_2, q_1), (q_2, q_2)\} & x \in (1,2) \\ \{(q_1, q_1), (q_2, q_2)\} & x \in (2,3). \end{cases}$$

Hence, $\mathcal{V}$ satisfies both conditions in Proposition 4.4.3.

In Example 4.4.2 we showed that $\mathcal{C}$ was equal to the HMM from Example 4.1.1. In Example 4.1.6 we showed that for the HMM from Example 4.1.1 the initial distributions e_{q_1} and e_{q_2} are distinguishable. By Proposition 4.4.3 it follows that the same initial distributions are distinguishable in the HMM $\mathcal{C}$.

Proof of Proposition 4.4.3. By Lemma 4.4.1, we have $d_{\mathcal{F}}(\pi_1, \pi_2) \leq d_{\mathcal{C}}(\pi_1, \pi_2)$. Hence $d_{\mathcal{F}}(\pi_1, \pi_2) = 1$ implies that $d_{\mathcal{C}}(\pi_1, \pi_2) = 1$.

Now assume $d_{\mathcal{F}}(\pi_1, \pi_2) < 1$. By the third point of Lemma 4.4.4 there is a word $w = a_1 \cdots a_n \in \tilde{\Sigma}^*$ and subdistributions μ_1, μ_2 such that $\mu_1 \leq \pi_1 M(w)$ and $\mu_2 \leq \pi_2 M(w)$ and $\mu_1 \equiv_{\mathcal{F}} \mu_2$ and $\|\mu_1\| = \|\mu_2\|$.

Fix $i \in [n]$. By the definition of M there is $V_i \in \mathcal{V}$ such that $M(a_i) = \int_{V_i} \Psi d\lambda$. Since $\mu_1 \leq \pi_1 M(w)$ we have that $M(a_i) \neq 0$ and hence $\lambda(V_i) > 0$. Define $\xi : \Sigma \to [0, \infty)$ as $\xi(x) = \min\{\Psi(x)_{q,r} \mid q, r \in Q, \Psi(x)_{q,r} > 0\}$. Write $\chi_{\xi \geq \delta} : \Sigma \to \{0,1\}$ for the function

$$\chi_{\xi \geq \delta}(x) = \begin{cases} 1 & \xi(x) \geq \delta \\ 0 & \xi(x) < \delta. \end{cases}$$

We have that $\chi_{\xi \geq \delta}\Psi$ converges monotonically to Ψ as $\delta \to 0$. Therefore, by the monotone convergence theorem $\lim_{\delta \to 0} \int_{V_i} \Psi \chi_{\xi \geq \delta} d\lambda = \int_{V_i} \Psi d\lambda$. It follows that there exists $\delta_i > 0$ such that $\lambda(\{x \in V_i \mid \xi(x) \geq \delta_i\}) > 0$.

By property 2 of $\mathcal{V}$ stated in the proposition, we have for all $x \in V_i$

$$\operatorname{supp}(\Psi(x)) = \operatorname{supp}\left(\int_{V_i} \Psi d\lambda\right) = \operatorname{supp}(M(a_i)) \tag{4.7}$$

since $\{\text{supp}\,(\Psi(x)) \mid x \in V_i\}$ contains at most one element.

We now relax the dependence on i. Let $\delta_{\min} = \min\{\delta_i \mid i \in [n]\}$. For each $i \in [n]$ we define $E_i := \{x \in V_i \mid \xi(x) \geq \delta_{\min}\}$ where it follows that $\lambda(E_i) > 0$. For each word $u \in \overline{E} := E_1 \times \cdots \times E_n$ and $j = 1, 2$ we have $\text{supp}\,(\pi_j \Psi(u)) = \text{supp}\,(\pi_j M(u))$ by (4.7). Further, $p_{min}\delta^n \mu_j \leq \pi_j \Psi(u)$ where p_{min} is the minimum non-zero entry of π_1 and π_2.

By Hahn's decomposition theorem there exists a set of infinite words $W \in \Sigma^\omega$ such that $d_{\mathcal{C}}(\pi_1, \pi_2) = \mathbb{P}_{\pi_1}(W) - \mathbb{P}_{\pi_2}(W)$. Let

$$
\begin{aligned}
X &= \{w' \in \Sigma^\omega \mid \exists\, u \in \overline{E} : uw' \in W\} \text{ and} \\
Y &= \{w' \in \Sigma^\omega \mid \exists\, u \in \overline{E} : uw' \in W^c\}.
\end{aligned}
$$

Suppose $w' \notin X$. Then, for all $u \in \overline{E}$ we have $uw' \notin W$ and therefore $uw' \in W^c$. Hence, $w' \in Y$ and so $X \cup Y = \Sigma^\omega$ by symmetry.

By property 1 of $\mathcal{V}$ stated in the proposition we have $\text{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\} = \text{span}\{M(a) \mid a \in \tilde{\Sigma}\}$. Since $\mu_1 \equiv_{\mathcal{F}} \mu_2$ we have $\mu_1 \equiv_{\mathcal{C}} \mu_2$ by Lemma 3.2.2. Thus,

$$
\begin{aligned}
1 - d_{\mathcal{C}}(\pi_1, \pi_2) &= 1 - \mathbb{P}_{\pi_1}(W) + \mathbb{P}_{\pi_2}(W) \\
&= \mathbb{P}_{\pi_1}(W^c) + \mathbb{P}_{\pi_2}(W) \\
&\geq \mathbb{P}_{\pi_1}(\overline{E}Y) + \mathbb{P}_{\pi_2}(\overline{E}X) \\
&\geq p_{min}\delta^n \Big(\mathbb{P}_{\mu_1}(Y) + \mathbb{P}_{\mu_2}(X)\Big) \\
&\geq p_{min}\delta^n \Big(\mathbb{P}_{\mu_1}(Y) + \mathbb{P}_{\mu_1}(X)\Big) \text{ since } \mu_1 \equiv_{\mathcal{C}} \mu_2 \\
&\geq p_{min}\delta^n \Big(\mathbb{P}_{\mu_1}(Y \cup X)\Big) \\
&= p_{min}\delta^n \Big(\mathbb{P}_{\mu_1}(\Sigma^\omega)\Big) \\
&> 0.
\end{aligned}
$$

□

4.4.1 Proof of Theorem 4.2.7

For the proof of Theorem 4.2.7 we use the following corollary to Lemma 3.4.1.

Corollary 4.4.6. *Let $\mathcal{C} = (Q, \Sigma, \Psi)$ and assume Σ has measure space $(\Sigma, \mathscr{G}, \lambda)$. We have that*

$$\text{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\} = \text{span}\left\{ \int_E \Psi d\lambda \mid E \in \mathscr{G} \right\}.$$

Proof. By Lemma 3.4.1 with $n = 1$, we have for any initial distributions π_1 and π_2 that $(\pi_1 - \pi_2)\Psi(w)\overline{1}^\top = 0$ for all $x \in \mathring{\Sigma}$ if and only if $(\pi_1 - \pi_2)\int_E \Psi \, d\lambda^n \overline{1}^\top = 0$ for all $E \in \mathscr{G}$. Hence

$$\begin{aligned}
\pi_1 - \pi_2 \perp \operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\} &\iff (\pi_1 - \pi_2)\Psi(w)\overline{1}^\top = 0 \quad \forall x \in \mathring{\Sigma} \\
&\iff (\pi_1 - \pi_2)\int_E \Psi \, d\lambda^n \overline{1}^\top = 0 \quad \forall E \in \mathscr{G} \\
&\iff \pi_1 - \pi_2 \perp \operatorname{span}\left\{ \int_E \Psi d\lambda \mid E \in \mathscr{G} \right\}. \qquad \square
\end{aligned}$$

We now restate and prove the main result of this section.

Theorem 4.2.7. *Let (Q, Σ, Ψ) be a continuous-observation HMM and let π_1, π_2 be initial distributions. For each $r_1 \in Q$ we define the projection*

$$\tilde{R}^{\pi_1,\pi_2}_{r_1} := \{r_2 \in Q \mid (r_1, r_2) \in \tilde{R}^{\pi_1,\pi_2}\}.$$

Then $d(\pi_1, \pi_2) < 1$ holds if and only if there are $r_1 \in Q$ and subdistributions μ_1, μ_2 such that

$$\mu_1 \equiv \mu_2 \text{ and } r_1 \in \mathit{supp}\,(\mu_1) \text{ and } \mathit{supp}\,(\mu_2) \subseteq \tilde{R}^{\pi_1,\pi_2}_{r_1}. \tag{4.5}$$

Proof. There is a finite set $\mathcal{E} \subseteq \mathscr{G}$ such that

$$\operatorname{span}\left\{ \int_E \Psi d\lambda \mid E \in \mathcal{E} \right\} = \operatorname{span}\left\{ \int_E \Psi d\lambda \mid E \in \mathscr{G} \right\}. \tag{4.8}$$

For each subset $A \subseteq Q \times Q$ we define a subset $S_A \subseteq \Sigma$ as

$$S_A = \{x \in \Sigma \mid \operatorname{supp}(\Psi(x)) = A\}.$$

Let $\mathcal{V}_{\text{supp}} = \{S_A \mid A \subseteq Q \times Q,\ \lambda(S_A) > 0\}$ and let $\mathcal{V}$ be the set of atomic elements of the finite σ-algebra $\sigma\{\mathcal{V}_{\text{supp}} \cup \mathcal{E}\}$. We show that $\mathcal{V}$ satisfies the properties in Proposition 4.4.3.

Let $E \in \mathcal{E}$ then there are $V_1, \cdots, V_n \in \mathcal{V}$ such that $E = V_1 \cup \cdots \cup V_n$. Therefore,

$$\int_E \Psi d\lambda = \int_{V_1 \cup \cdots \cup V_n} \Psi d\lambda = \sum_{i=1}^n \int_{V_i} \Psi d\lambda$$

and so

$$\begin{aligned}
\operatorname{span}\left\{ \int_E \Psi d\lambda \mid E \in \mathcal{E} \right\} &\subseteq \operatorname{span}\left\{ \int_V \Psi d\lambda \mid V \in \mathcal{V} \right\} \\
&\subseteq \operatorname{span}\left\{ \int_E \Psi d\lambda \mid E \in \mathscr{G} \right\} \\
&= \operatorname{span}\{\Psi(x) \mid x \in \mathring{\Sigma}\} \text{ by Corollary 4.4.6.}
\end{aligned} \tag{4.9}$$

By property (4.8) we actually have equality in (4.9).

Let $V \in \mathcal{V}$ then since $\mathcal{V}_{\text{supp}} \subseteq \sigma\{\mathcal{V}\}$ there is a $V' \in \mathcal{V}_{\text{supp}}$ such that $V \subseteq V'$. Since

$$\operatorname{supp}\Psi(V) = \operatorname{supp}\Psi(V')$$

it follows that the image $\operatorname{supp}\Psi(V)$ contains exactly one element.

Therefore, by Proposition 4.4.3 we have $d_{\mathcal{C}}(\pi_1, \pi_2) < 1$ if and only if $d_{\mathcal{F}}(\pi_1, \pi_2) < 1$ where $\mathcal{F}$ is the $\mathcal{V}$-approximation of $\mathcal{C}$. Suppose $\mathcal{F}$ has joint-reachability set R^{π_1,π_2}. Recall that for each $r_1 \in Q$ we have

$$R^{\pi_1,\pi_2}_{r_1} := \{r_2 \in Q \mid (r_1, r_2) \in R^{\pi_1,\pi_2}\}.$$

By Theorem 4.1.5 we have $d_{\mathcal{F}}(\pi_1, \pi_2) < 1$ if and only if there is $r_1 \in Q$ and subdistributions μ_1, μ_2 such that

$$\mu_1 \equiv_{\mathcal{F}} \mu_2 \text{ and } r_1 \in \operatorname{supp}(\mu_1) \text{ and } \operatorname{supp}(\mu_2) \subseteq R^{\pi_1,\pi_2}_{r_1}. \tag{4.10}$$

Since (4.9) holds, by Lemma 3.2.2, we have for any two subdistributions μ_1, μ_2 that $\mu_1 \equiv_{\mathcal{C}} \mu_2$ if and only if $\mu_1 \equiv_{\mathcal{F}} \mu_2$. For all $V \in \mathcal{V}$ we have $\{\operatorname{supp}(\Psi(x)) \mid x \in V\}$ contains at most one element and so $\operatorname{supp}\left(\int_V \Psi d\lambda\right) = \operatorname{supp}(\Psi(x))$ for all $x \in V$. It follows that $G_{\mathcal{F},\mathcal{F}} = \tilde{G}_{\mathcal{C},\mathcal{C}}$.

By Lemma 4.1.3 and Lemma 4.2.6 we have

$$\begin{aligned}
R^{\pi_1,\pi_2} &= \Big\{(r_1, r_2) \in Q \times Q \;\Big|\; \operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2) \to_{G_{\mathcal{F},\mathcal{F}}} \{(r_1, r_2)\}\Big\} \\
&= \Big\{(r_1, r_2) \in Q \times Q \;\Big|\; \operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2) \to_{\tilde{G}_{\mathcal{C},\mathcal{C}}} \{(r_1, r_2)\}\Big\} \\
&= \tilde{R}^{\pi_1,\pi_2}.
\end{aligned}$$

Therefore, the condition (4.10) is equivalent to

$$\mu_1 \equiv_{\mathcal{C}} \mu_2 \text{ and } r_1 \in \operatorname{supp}(\mu_1) \text{ and } \operatorname{supp}(\mu_2) \subseteq \tilde{R}^{\pi_1,\pi_2}_{r_1}. \qquad \square$$

5
Profile Languages

In Chapter 3 and Chapter 4 we described sufficient conditions on the encoding of an observation density matrix for decidability of equivalence and distinguishability. Specifically the following

1. In Theorem 3.3.1 we showed that if a HMM is over a linearly decomposable profile language then we may decide in polynomial time whether two initial distributions are equivalent.

2. In Theorem 4.2.9 we showed that if a HMM is over a linearly decomposable and joint graph decomposable profile language then we may decide in polynomial time whether two initial distributions are distinguishable.

In this chapter, we give sufficient conditions for a profile language to be linearly decomposable and joint graph decomposable. We aim for generality in these conditions and show that the profile language Γ_{GEM} introduced in Section 3.3 and moreover a much larger profile language (which contains Γ_{GEM}) also has these properties. The techniques presented here may be generally applicable to even larger profile languages.

In Section 5.2, we introduce the Span-$\mathcal{I}$ operator on profile languages and show that if a profile language is linearly decomposable then applying the Span-$\mathcal{I}$ operator preserves this property. We also show that if a profile language encodes a set of analytic functions that is linearly independent, then the Span-$\mathcal{I}$ operator produces a profile language that is joint graph decomposable.

In Section 5.3 we describe one general method for showing that probability density functions are linearly independent and thus any profile language that encodes them is linearly decomposable.

Finally, in Section 5.4 we show that the set of observations Σ need not be one-dimensional. We consider a profile language Γ_D^k which encodes Dirichlet distributions and show that Γ_D^k is linearly decomposable and joint graph decomposable.

5.1 Profile Languages from Analytic Functions

Recall the following:

Definition 3.3.1. Let Γ be a profile language with $[\![\Gamma]\!] \subseteq \mathcal{L}_1(\Sigma, \lambda)$. We say that Γ is *linearly decomposable* if there is a procedure that takes a finite subset $\{\gamma_1, \ldots, \gamma_n\} = \Gamma_0 \subseteq \Gamma$ and produces profiles $\beta_1, \ldots, \beta_m \in \Gamma_0$ and a set of rational coefficients $b_{i,j} \in \mathbb{Q}$ for $i \in [n], j \in [m]$ such that $\{[\![\beta_1]\!], \ldots, [\![\beta_m]\!]\}$ is a basis for $\text{span}\{[\![\gamma_1]\!], \ldots, [\![\gamma_n]\!]\}$ (hence $m \leq n$), and

$$[\![\gamma_i]\!] =_\lambda \sum_{j=1}^{m} b_{i,j}[\![\beta_j]\!] \text{ for all } i \in [n]. \tag{3.2}$$

Further, this procedure takes polynomial time in the total length of the encoding of Γ_0.

Definition 4.2.2. Let Γ be a profile language with $[\![\Gamma]\!] \subseteq \mathcal{L}_1(\Sigma, \lambda)$. We say that Γ is joint graph decomposable if there is a procedure that takes a pair of profiles $\alpha, \beta \in \Gamma$ and decides whether $\lambda(\text{supp}\,([\![\alpha]\!]) \cap \text{supp}\,([\![\beta]\!])) > 0$ and this procedure takes polynomial time in the combined length of the encodings of α and β.

We will establish sufficient conditions for a profile language to be linearly decomposable and joint graph decomposable. In particular these conditions will be satisfied by a profile language encoding a large class of standard probability distribution functions.

Let $\mathbb{I} = \{(a,b) \subseteq \mathbb{R} \mid a, b \in \mathbb{Q} \cup \{-\infty, \infty\} : a < b\}$ and let γ be a profile. We define a new profile $\mathcal{I}(\gamma, a, b)$ obtained by restricting the support of $[\![\gamma]\!]$ to an interval $(a,b) \in \mathbb{I}$. More concretely, the profile $\mathcal{I}(\gamma, a, b)$ encodes the function $[\![\gamma]\!]\chi_{(a,b)}$. Let

Γ be a profile language, then Span-$\mathcal{I}(\Gamma)$ is the profile language obtained by closing Γ under interval-restriction and then rational linear combinations. That is,

$$\begin{aligned}\text{Span-}\mathcal{I}(\Gamma) = \{\mathcal{S}(\lambda_1, \mathcal{I}(\gamma_1, a_1, b_1)), \dots, \lambda_n, \mathcal{I}(\gamma_n, a_n, b_n)) \mid & \lambda_1, \dots, \lambda_n \in \mathbb{Q} \setminus \{0\}, \\ & \gamma_1, \dots, \gamma_n \in \Gamma, \\ & (a_1, b_1), \dots, (a_n, b_n) \in \mathbb{I}\}.\end{aligned}$$

Example 5.1.1. Suppose γ encodes the monomial function $x \mapsto x^k$ for $k \in \mathbb{N}_0$. Then, $\mathcal{I}(\gamma, a, b)$ encodes the function $x \mapsto x^k \chi_{(a,b)}(x)$ where $a, b \in \mathbb{Q}$ and $a < b$. We have that $\mathcal{S}(1, \mathcal{I}(\gamma, a, b))$ also encodes $x \mapsto x^k \chi_{(a,b)}(x)$.

Let Γ_m be the set of profiles encoding monomial functions. Hence, $[\![\Gamma_m]\!] = \{x \mapsto x^k \mid k \in \mathbb{N}_0\}$. We have that

$$[\![\text{Span-}\mathcal{I}(\Gamma_m)]\!] = \Big\{x \mapsto \sum_{(\lambda,i,(a,b)) \in A} \lambda x^i \chi_{(a,b)}(x) \mid \text{finite } A \subseteq (\mathbb{Q} \setminus \{0\}) \times (\mathbb{N}_0) \times \mathbb{I}\Big\}.$$

It follows that $\mathcal{S}(1, \mathcal{I}(\gamma, a, b)) \in$ Span-$\mathcal{I}(\Gamma_m)$ and thus any interval-domain monomial can be encoded by an element of Span-$\mathcal{I}(\Gamma_m)$.

Let $\gamma_i \in \Gamma_m$ encode the monomial $x \mapsto x^i$ for $i \in \mathbb{N}_0$. Then for $n \in \mathbb{N}_0$ we have

$$\mathcal{S}(\lambda_0, \mathcal{I}(\gamma_0, -\infty, \infty), \dots, \lambda_n, \mathcal{I}(\gamma_n, -\infty, \infty))$$

encodes the polynomial $x \mapsto \sum_{i=0}^{n} \lambda_i x^i$. It follows that any polynomial with rational coefficients can be encoded by an element of Span-$\mathcal{I}(\Gamma_m)$.

Definition 5.1.1. For any open set $D \subseteq \mathbb{R}$, we say a function $f : D \to \mathbb{R}$ is *analytic* on D if for any $x_0 \in D$ one can write

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n \tag{5.1}$$

where the coefficients $a_0, a_1, \dots$ are real numbers and the series (5.1) is convergent to $f(x)$ for x in a neighbourhood of x_0.

Suppose there is an analytic function $\overline{f} : \mathbb{R} \to [0, \infty)$ such that $\overline{f}(x) = f(x)$ for all $x \in D$. We call $\overline{f}$ an analytic continuation of f to $\mathbb{R}$. The following is proven in Appendix A.3.

Lemma 5.1.2. *Let $D \subseteq \mathbb{R}$ and let $f : D \to \mathbb{R}$ be analytic on D. If an analytic continuation of f to $\mathbb{R}$ exists, then it is unique.*

Example 5.1.3. Let $x_0 \in \mathbb{R}$ and let $a_0 = \frac{1}{2}x_0^2$, $a_1 = -x_0$ and $a_2 = -\frac{1}{2}$. Further, define $g : \mathbb{R} \to \mathbb{R}$ where $g(x) = -\frac{1}{2}x^2$. Then,

$$\frac{1}{2}x_0^2 - x_0(x - x_0) - \frac{1}{2}(x - x_0)^2 = -\frac{1}{2}x^2$$

Therefore, g is analytic on $\mathbb{R}$.

Define $h : \mathbb{R} \to \mathbb{R}$ where $h(x) = -\frac{1}{\sqrt{2\pi}}e^x$. Then let $b_n = \frac{e^{x_0}}{n!\sqrt{2\pi}}$ for all $n \in \mathbb{N}_0$ where we write $0! = 0$. It follows by the power series representation of the exponential function that

$$\begin{aligned}\sum_{n=0}^{\infty} b_n(x - x_0)^n &= \frac{e^{x_0}}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{1}{n!}(x - x_0)^n \\ &= \frac{e^{x_0}}{\sqrt{2\pi}} e^{x - x_0} \\ &= \frac{1}{\sqrt{2\pi}} e^x.\end{aligned}$$

Thus, h is analytic on $\mathbb{R}$.

We have $(h \circ g)(x)$ is the pdf of a gaussian distribution which is analytic on $\mathbb{R}$ since the composition of two analytic functions on $\mathbb{R}$ is also analytic (see Proposition 1.3.3 on page 17 of [34])

Table 5.1 gives a summary of some standard probability distributions. Each

Table 5.1: Some standard probability distribution functions

Distribution Name	pdf (in terms of x)	Domain	Parameters	Profile Language
Student's t	$\frac{\tilde{\Gamma}((\nu+1)2)}{\sqrt{\nu\pi}\tilde{\Gamma}(\nu/2)}(1 + \frac{x^2}{\nu})^{-(\nu+1)/2}$	$\mathbb{R}$	$\nu \in \mathbb{Q}_+$	Γ_t
Chi-squared	$\frac{1}{2^{k/2}\tilde{\Gamma}(k/2)}x^{k/2-1}\exp(-x/2)$	$(0, \infty)$	$k \in \mathbb{N} \setminus \{1\}$	Γ_χ
Exponential	$\lambda \exp(-\lambda x)$	$(0, \infty)$	$\lambda \in \mathbb{Q}_+$	Γ_e
Maxwell-Boltzmann	$\sqrt{\frac{2}{\pi}}\frac{1}{a^3}x^2 \exp(-x^2/(2a^2))$	$(0, \infty)$	$a \in \mathbb{Q}_+$	Γ_{mb}
Gaussian	$\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$	$\mathbb{R}$	$\mu \in \mathbb{Q}, \sigma \in \mathbb{Q}_+$	Γ_g

distribution has an associated pdf. In the pdf of the Chi-squared distribution $\tilde{\Gamma}$ is the gamma function where $\tilde{\Gamma}(n) = \int_0^\infty x^{n-1}e^{-x}dx$. Each distribution has parameters that control the precise expression of the pdf. For example, the mean μ and standard deviation σ are the parameters of the Gaussian distribution.

We introduce a profile language for the analytic continuation of each pdf in Table 5.1 so that each profile encodes a function with the same domain. For example $\Gamma_e = \{\mathcal{A}_e(\lambda) \mid \lambda \in \mathbb{Q}_+\}$ where we define $\mathcal{A}_e$ as the constructor for exponential functions and $[\![\mathcal{A}_e(\lambda)]\!] : \mathbb{R} \to [0, \infty)$ such that $[\![\mathcal{A}_e(\lambda)]\!](x) = \lambda \exp(-\lambda x)$. The pdf of an exponential distribution with rate λ has domain $(0, \infty)$ and so is not directly encodable with Γ_e. However, we may encode the pdf with the profile $\mathcal{I}(\mathcal{A}_e(\lambda), 0, \infty)$. Therefore, every exponential distribution pdf can be encoded by Span-$\mathcal{I}(\Gamma_e)$.

Later in this chapter we prove the following.

Proposition 5.1.4. *Let* Γ *be a linearly decomposable profile language encoding analytic functions on* $\mathbb{R}$*. Then* Span-$\mathcal{I}(\Gamma)$ *is linearly decomposable.*

Let Γ be a (not necessarily finite) profile language. We say $[\![\Gamma]\!]$ is linearly independent if for every finite subset $\Gamma_0 \subseteq \Gamma$ we have that the set of functions $[\![\Gamma_0]\!]$ is linearly independent in the vector space $\mathcal{L}_1(\Sigma, \lambda)$. Using this definition there is an analogue to Proposition 5.1.4 for joint graph decomposable profile languages.

Proposition 5.1.5. *Let* Γ *be a profile language encoding analytic functions on* $\mathbb{R}$*. Additionally, assume* $[\![\Gamma]\!]$ *is linearly independent. Then,* Span-$\mathcal{I}(\Gamma)$ *is joint graph decomposable.*

Recall Γ_m from Example 5.1.1 is the profile language encoding the set of monomial functions. We may extend the argument of Example 5.3.1 to a larger set of profiles with the following lemma.

Lemma 5.1.6. *The set of functions* $[\![\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g]\!]$ *is linearly independent.*

The following theorem is the main result of this section.

Theorem 5.1.7. *Let* Γ *be a profile language encoding analytic functions on* $\mathbb{R}$ *and assume* $[\![\Gamma]\!]$ *is linearly independent. Then* Span-$\mathcal{I}(\Gamma)$ *is linearly decomposable and joint graph decomposable.*

Proof. We have Span-$\mathcal{I}(\Gamma)$ is joint graph decomposable by Proposition 5.1.5. Since $[\![\Gamma]\!]$ is linearly independent, for any finite subset $\Gamma_0 \subseteq \Gamma$ we have $[\![\Gamma_0]\!]$ is linearly independent. Thus Γ is trivially linearly decomposable. Therefore by Proposition 5.1.4 we have Span-$\mathcal{I}(\Gamma)$ is linearly decomposable. □

This immediately implies the following.

Corollary 5.1.8. *The profile language* Span-$\mathcal{I}(\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g)$ *is linearly decomposable and joint graph decomposable.*

Recall the profile language Γ_{GEM} introduced in Section 3.3. Using Corollary 5.1.8, we prove the following at the end of Section 5.3.

Proposition 5.1.9. Γ_{GEM} *is linearly decomposable and joint graph decomposable.*

5.2 Decomposability of Span-$\mathcal{I}$

In this section we prove Proposition 5.1.4 and Proposition 5.1.5. We use the *Identity Theorem* to establish that linear dependence on an open interval implies linear dependence on $\mathbb{R}$ in Lemma 5.2.3.

Let $S \subseteq \mathbb{R}$. We say $x \in S$ is an *accumulation point* of S if there exists a sequence $(x_n)_{n=1}^{\infty}$ such that $x_n \in S$ and $x_n \neq x$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} x_n = x$.

Theorem 5.2.1. *[Identity Theorem] Let $f : \mathbb{R} \to \mathbb{R}$ be analytic on $\mathbb{R}$. If $f(x) = 0$ for all $x \in S \subseteq \mathbb{R}$, where S has an accumulation point, then $f(x) = 0$ for all $x \in \mathbb{R}$.*

Specifically, we use the following a corollary to the Identity Theorem proven in Appendix A.3.

Corollary 5.2.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be analytic on $\mathbb{R}$. If $f(x) = 0$ for all $x \in S \subseteq \mathbb{R}$, where $\lambda_{Leb}(S) > 0$, then $f(x) = 0$ for all $x \in \mathbb{R}$.*

Lemma 5.2.3. *Suppose $f_1, \ldots, f_n : \mathbb{R} \to [0, \infty)$ are analytic on $\mathbb{R}$ and $\{f_1, \ldots, f_n\}$ is linearly independent. Then for any set $A \subseteq \mathbb{R}$ such that $\lambda_{Leb}(A) > 0$, we have that $\{\chi_A f_1, \ldots, \chi_A f_n\}$ is linearly independent.*

Proof. Suppose there is a linear dependence

$$\sum_{i=1}^{n} \beta_i \chi_A f_i = 0$$

Then since $\sum_{i=1}^{n} \beta_i f_i$ is analytic and $\sum_{i=1}^{n} \beta_i f_i(x) = 0$ for all $x \in A$, by the Identity Theorem (Theorem 5.2.1) we have $\sum_{i=1}^{n} \beta_i f_i(x) = 0$ for all $x \in \mathbb{R}$. Since $\{f_1, \ldots, f_n\}$ is linearly independent it follows that $\beta_1 = \cdots = \beta_n = 0$. □

Proposition 5.1.4. *Let* Γ *be a linearly decomposable profile language encoding analytic functions on* $\mathbb{R}$. *Then* Span-$\mathcal{I}(\Gamma)$ *is linearly decomposable.*

Proof. Consider a finite set of profiles $\Gamma_0 \subseteq$ Span-$\mathcal{I}(\Gamma)$. For each $\gamma \in \Gamma_0$ we have

$$\gamma = \mathcal{S}(\lambda_1, \mathcal{I}(\gamma_1, a_1, b_1), \ldots, \lambda_N, \mathcal{I}(\gamma_N, a_N, b_N)) \tag{5.2}$$

for some $N \in \mathbb{N}$, $\gamma_1, \ldots, \gamma_N \in \Gamma$, $\lambda_1, \ldots, \lambda_N \in \mathbb{Q} \setminus \{0\}$ and $(a_1, b_1), \ldots, (a_N, b_N) \in \mathbb{I}$.

Collect together in $I_0 \subseteq \mathbb{I}$ all the intervals that appear in the description of at least one profile in Γ_0. By sorting the start and end points of intervals in I_0 we may compute a new finite set $I \subseteq \mathbb{I}$ such that the intervals in I are disjoint and every interval in I_0 is the union of intervals in I up to a λ_{Leb}-null set.

Collect together into Γ_1 all the profiles in Γ that appear in the description of at least one profile in Γ_0. Since $\Gamma_1 \subseteq \Gamma$ is finite and Γ is linearly decomposable we may compute $\Gamma_2 \subseteq \Gamma_1$ such that $[\![\Gamma_2]\!]$ is linearly independent.

Let $\Gamma_3 = \{\mathcal{I}(\gamma, a, b) \mid \gamma \in \Gamma_2, (a, b) \in I\}$. We show that $[\![\Gamma_3]\!]$ is linearly independent. Suppose there is a linear dependence $\sum_{(a,b)\in I} \sum_{\gamma\in\Gamma_2} c_{a,b,\gamma} \chi_{(a,b)}(x)[\![\gamma]\!](x) = 0$ for all $x \in \mathbb{R}$. Fix $(a', b') \in I$. For all $x \in (a', b')$ we have

$$0 = \sum_{(a,b)\in I} \sum_{\gamma\in\Gamma_2} c_{a,b,\gamma} \chi_{(a,b)}(x)[\![\gamma]\!](x) = \sum_{\gamma\in\Gamma_2} c_{a',b',\gamma} \chi_{(a',b')}(x)[\![\gamma]\!](x).$$

By Lemma 5.2.3 the set of functions $\{\chi_{(a',b')}[\![\gamma]\!] \mid \gamma \in \Gamma_2\}$ is linearly independent and so $c_{\gamma,a',b'} = 0$ for all $\gamma \in \Gamma_2$. We can repeat this argument for all intervals in I to show that $[\![\Gamma_3]\!]$ is linearly independent.

We may compute the coordinates of all functions in $[\![\Gamma_1]\!]$ in terms of the set of functions $[\![\Gamma_2]\!]$ since Γ is linearly decomposable. Hence, we may compute coordinates for $[\![\mathcal{I}(\gamma, a, b)]\!]$ where $\gamma \in \Gamma_1$ and $(a, b) \in I_0$ in terms of the set of functions $[\![\Gamma_3]\!]$. Each profile in Γ_0 has the form given in (5.2) and therefore since $[\![\Gamma_3]\!]$ is linearly independent, we may compute (unique) coordinates of all functions in $[\![\Gamma_0]\!]$ in terms of the set of functions $[\![\Gamma_3]\!]$.

With these final set of coordinates at hand, we compute a subset $\mathcal{B} \subseteq \Gamma_0$ such that $[\![\mathcal{B}]\!]$ is a basis of span $[\![\Gamma_0]\!]$ as follows. Starting with the $\mathcal{B} = \emptyset$, go through Γ_0 one by one; whenever a profile $\gamma \in \Gamma_0$ is such that $[\![\mathcal{B} \cup \{\gamma\}]\!]$ is linearly independent then add γ to $\mathcal{B}$. The check for linear independence can be performed in terms of the computed coordinates of $[\![\Gamma_0]\!]$ in the basis $[\![\Gamma_3]\!]$. For the final set $\mathcal{B}$ we have that $[\![\mathcal{B}]\!]$ is a basis of span$[\![\Gamma_0]\!]$. The coefficients that express $[\![\Gamma_0]\!]$ as a linear combination of $[\![\mathcal{B}]\!]$ can be computed similarly. The computation described in this proof takes time polynomial in the size of the input encoding. $\square$

Lemma 5.2.4. *Let $f_1, \dots, f_n : \mathbb{R} \to \mathbb{R}$ be analytic on $\mathbb{R}$, let $\lambda_1, \dots, \lambda_n \in \mathbb{Q} \setminus \{0\}$ and let $(a_1, b_1), \dots, (a_n, b_n) \in \mathbb{I}$. Assume that $\{f_1, \dots, f_n\}$ is linearly independent then,*

$$\mathit{supp}\left(\sum_{k=1}^{n} \lambda_k f_k \chi_{(a_k,b_k)}\right) = \bigcup_{k=1}^{n} (a_k, b_k)$$

Proof. Clearly $\operatorname{supp}(\sum_{k=1}^{n} \lambda_k f_k \chi_{(a_k,b_k)}) \subseteq \bigcup_{k=1}^{n}(a_k, b_k)$. Suppose there is $C \subseteq \bigcup_{k=1}^{n}(a_k, b_k)$ such that $\lambda_{Leb}(C) > 0$ and $\sum_{k=1}^{n} \lambda_k f_k(x)\chi_{(a_k,b_k)}(x) = 0$ for all $x \in C$. Without loss of generality assume there is $1 \le n_0 \le n$ such that $\lambda_{Leb}((a_k, b_k) \cap C) > 0$ for all $k \in [n_0]$. Then, $\sum_{k=1}^{n_0} \lambda_k f_k(x) = 0$ for all $x \in C \cap \bigcap_{k=1}^{n_0}(a_k, b_k)$. Since $\{f_1, \dots, f_n\}$ is linearly independent it follows that $\lambda_1 = \cdots = \lambda_{n_0} = 0$ which contradicts the definition of $\lambda_1, \dots, \lambda_n$. $\square$

Proposition 5.1.5. *Let Γ be a profile language encoding analytic functions on $\mathbb{R}$. Additionally, assume $[\![\Gamma]\!]$ is linearly independent. Then,* Span-$\mathcal{I}(\Gamma)$ *is joint graph decomposable.*

Proof. Let $\gamma_1, \gamma_2 \in$ Span-$\mathcal{I}(\Gamma)$. We have for each $i = 1, 2$

$$\gamma_i = S(\lambda_{i,1}, \mathcal{I}(\gamma_{i,1}, (a_{i,1}, b_{i,2})), \dots, \lambda_{i,N_i}, \mathcal{I}(\gamma_{i,N_i}, (a_{i,N_i}, b_{i,N_i})))$$

for some $N_i \in \mathbb{N}$, $\gamma_{i,1}, \dots, \gamma_{i,N_i} \in \Gamma$, $\lambda_{i,1}, \dots, \lambda_{i,N_i} \in \mathbb{Q}\setminus\{0\}$ and $(a_{i,1}, b_{i,1}), \dots, (a_{i,N_i}, b_{i,N_i}) \in$

$\mathbb{I}$. By Lemma 5.2.4 we have $\operatorname{supp}(\llbracket\gamma_i\rrbracket) = \bigcup_{k=1}^{N_i}(a_{i,k}, b_{i,k})$ for $i = 1, 2$. Therefore

$$\begin{aligned}\operatorname{supp}(\llbracket\gamma_1\rrbracket) \cap \operatorname{supp}(\llbracket\gamma_2\rrbracket) &= \bigcup_{k=1}^{N_1}(a_{1,k}, b_{1,k}) \cap \bigcup_{l=1}^{N_2}(a_{2,l}, b_{2,l}) \\ &= \bigcup_{k=1}^{N_1}\bigcup_{l=1}^{N_2}(a_{1,k}, b_{1,k}) \cap (a_{2,l}, b_{2,l}) \\ &= \bigcup_{k=1}^{N_1}\bigcup_{l=1}^{N_2}(\max\{a_{1,k}, a_{2,l}\}, \min\{b_{1,k}, b_{2,l}\})\end{aligned}$$

where we use the notation $(\max\{a_{1,k}, a_{2,l}\}, \min\{b_{1,k}, b_{2,l}\}) = \emptyset$ if $\max\{a_{1,k}, a_{2,l}\} \geq \min\{b_{1,k}, b_{2,l}\}$. Thus we can decide whether $\lambda_{Leb}(\operatorname{supp}(\llbracket\gamma_1\rrbracket) \cap \operatorname{supp}(\llbracket\gamma_2\rrbracket)) > 0$ by checking whether there exists $k \in [N_1]$ and $l \in [N_2]$ such that $\max\{a_{1,k}, a_{2,l}\} < \min\{b_{1,k}, b_{2,l}\}$. This procedure takes polynomial time. $\square$

5.3 Asymptotic Ordering

In this section we define *asymptotic ordering* which we use to establish linear independence of functions with domain $\mathbb{R}$. In particular we prove Lemma 5.1.6. The techniques presented in this section are generally applicable and can be used to prove similar results for even larger profile languages. However, there may be other techniques. For example, another general technique to show a set of functions is linearly independent is to examine their Wronskian matrix [35]. We end this section by proving Proposition 5.1.9 and thus tying up the loose ends from Chapter 3 and Chapter 4: Corollary 3.3.5 and Corollary 4.2.10 respectively.

Definition 5.3.1. Let $g, h : \mathbb{R} \to \mathbb{R}$ and write $g \prec h$ if for all $\varepsilon > 0$ there exists $M \in \mathbb{N}$ such that $|g(x)| < \varepsilon|h(x)|$ for all $x \geq M$. The relation $\prec$ is transitive. The notation $f \prec h$ is equivalent to $f = o(h)$ where o refers to little-o notation[36]. A profile language Γ is *asymptotically ordered* if for every $\gamma_1, \gamma_2 \in \Gamma$ such that

$$\lambda\Big(\Big\{x \in \mathbb{R} \;\Big|\; \llbracket\gamma_1\rrbracket(x) \neq \llbracket\gamma_2\rrbracket(x)\Big\}\Big) > 0$$

either $\llbracket\gamma_1\rrbracket \prec \llbracket\gamma_2\rrbracket$ or $\llbracket\gamma_2\rrbracket \prec \llbracket\gamma_1\rrbracket$.

Example 5.3.1. Let $\lambda_1, \lambda_2 \in \mathbb{Q}_+$ with $\lambda_1 > \lambda_2$ and let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ with $f_i(x) = \lambda_i \exp(-\lambda_i x)$ for $i = 1, 2$. It follows that $f_1(x)/f_2(x) < 1$ for all $x > 0$ and $f_1(x)/f_2(x) = \frac{\lambda_1}{\lambda_2}\exp(-(\lambda_1 - \lambda_2)x) \to 0$ as $x \to \infty$. Therefore $\{x \in \mathbb{R} \mid f_1(x) < f_2(x)\} = \mathbb{R}$ and $f_1 \prec f_2$. It follows that Γ_e is asymptotically ordered.

Lemma 5.3.2. *Suppose* Γ *is an asymptotically ordered profile language then* $[\![\Gamma]\!]$ *is linearly independent.*

Proof. Let $\{\gamma_1, \ldots, \gamma_n\} \subseteq \Gamma$. Since Γ is asymptotically ordered without loss of generality we may assume $[\![\gamma_1]\!] \prec \cdots \prec [\![\gamma_n]\!]$. Suppose there is a linear dependence such that

$$\sum_{i=1}^{n} \beta_i [\![\gamma_i]\!](x) = 0$$

for all $x \in \mathbb{R}$. Then since $[\![\gamma_1]\!] \prec \cdots \prec [\![\gamma_n]\!]$, by transitivity we have that $[\![\gamma_i]\!] \prec [\![\gamma_n]\!]$ for all $i \in [n-1]$. Therefore, for all $i \in [n-1]$ there exists $M_i \in \mathbb{N}$ such that $|[\![\gamma_i]\!](x)| < \varepsilon [\![\gamma_n]\!](x)$ for all $x \geq M_i$. Let $M = \max\{M_1, \ldots, M_{n-1}\}$. For any $x \geq M$ we have $0 \leq |[\![\gamma_{n-1}]\!](x)| < \varepsilon [\![\gamma_n]\!](x)$ hence $[\![\gamma_n]\!](x) > 0$ and also,

$$\begin{aligned}
|\beta_n| &= \frac{1}{[\![\gamma_n]\!](x)} |\beta_n [\![\gamma_n]\!](x)| \\
&= \frac{1}{[\![\gamma_n]\!](x)} |\sum_{i=1}^{n-1} \beta_i [\![\gamma_i]\!](x)| \\
&\leq \frac{1}{[\![\gamma_n]\!](x)} \sum_{i=1}^{n-1} |\beta_i| |[\![\gamma_i]\!](x)| \\
&< \frac{1}{[\![\gamma_n]\!](x)} \Big(\sum_{i=1}^{n-1} |\beta_i| \varepsilon [\![\gamma_n]\!](x) \Big) \\
&= \varepsilon \Big(\sum_{i=1}^{n-1} |\beta_i| \Big).
\end{aligned}$$

By letting $\varepsilon \to 0$ we have that $\beta_n = 0$. We may repeat this argument for decreasing n to establish that $\beta_1 = \cdots = \beta_n = 0$. Therefore, $[\![\{\gamma_1, \ldots, \gamma_n\}]\!]$ is linearly independent. □

Using the previous lemma we may now prove the following.

Lemma 5.1.6. *The set of functions* $[\![\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g]\!]$ *is linearly independent.*

Lemma 5.1.6 is immediate from Lemma 5.3.2 and the next result.

Lemma 5.3.3. *The profile language* $\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g$ *is asymptotically ordered.*

Proof. Let $\gamma_1, \gamma_2 \in \Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g$ and $\lambda(\{x \in \mathbb{R} \mid [\![\gamma_1]\!](x) \neq [\![\gamma_2]\!](x)\}) > 0$.

We consider first the following cases:

1. $\gamma_1, \gamma_2 \in \Gamma_m$,
2. $\gamma_1, \gamma_2 \in \Gamma_t$,
3. $\gamma_1, \gamma_2 \in \Gamma_\chi$,
4. $\gamma_1, \gamma_2 \in \Gamma_e$,
5. $\gamma_1, \gamma_2 \in \Gamma_{mb}$,
6. $\gamma_1, \gamma_2 \in \Gamma_g$.

In each of these cases, $\lambda(\{x \in \mathbb{R} \mid [\![\gamma_1]\!](x) \neq [\![\gamma_2]\!](x)\}) > 0$ is enough to establish that γ_1 and γ_2 differ on at least one of their parameters. We show that either $\lim_{x\to\infty}[\![\gamma_1]\!](x)/[\![\gamma_2]\!](x) = 0$ or $\lim_{x\to\infty}[\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) = 0$ which implies either $[\![\gamma_1]\!] \prec [\![\gamma_2]\!]$ or $[\![\gamma_2]\!] \prec [\![\gamma_1]\!]$.

Case 1. Suppose $[\![\gamma_i]\!](x) = x^{k_i}$ for $i = 1, 2$ with $k_1 < k_2$. We have $x^{k_1}/x^{k_2} = x^{k_1 - k_2} \to 0$ as $x \to \infty$.

Case 2. Suppose $[\![\gamma_i]\!](x) = \frac{\tilde{\Gamma}((\nu_i+1)/2)}{\sqrt{\nu_i\pi}\tilde{\Gamma}(\nu_i/2)}(1 + \frac{x^2}{\nu_i})^{-(\nu_i+1)/2}$ for $i = 1, 2$. We have

$$\lim_{x\to\infty} \frac{(1 + \frac{x^2}{\nu_i})^{-(\nu_i+1)/2}}{x^{-(\nu_i+1)}} = 1$$

for $i = 1, 2$. Assume $\nu_1 > \nu_2$ then

$$\lim_{x\to\infty}[\![\gamma_1]\!](x)/[\![\gamma_2]\!](x) = \lim_{x\to\infty} x^{-(\nu_1+1)}/x^{-(\nu_2+1)} = \lim_{x\to\infty} x^{\nu_2-\nu_1} = 0.$$

Case 3. Suppose $[\![\gamma_i]\!](x) = \frac{1}{2^{k_i/2}\tilde{\Gamma}(k_i/2)} x^{k_i/2-1} \exp -x/2$ for $i = 1, 2$. We have

$$\lim_{x\to\infty} \frac{\frac{1}{2^{k_i/2}\tilde{\Gamma}(k_i/2)} x^{k_i/2-1} \exp -x/2}{x^{k_i/2-1} \exp -x/2} = 0 \tag{5.3}$$

for $i = 1, 2$. Assume $k_1 < k_2$ then

$$\lim_{x\to\infty}[\![\gamma_1]\!](x)/[\![\gamma_2]\!](x) = \lim_{x\to\infty} \frac{x^{k_1/2-1}}{x^{k_2/2-1}} = \lim_{x\to\infty} x^{(k_1-k_2)/2} = 0.$$

Case 4. Suppose $[\![\gamma_i]\!](x) = \lambda_i \exp -\lambda_i x$ for $i = 1, 2$ where $\lambda_1 > \lambda_2$. We have

$$\lim_{x\to\infty} [\![\gamma_1]\!](x)/[\![\gamma_2]\!](x) = \lim_{x\to\infty} \frac{\lambda_1}{\lambda_2} \exp -(\lambda_1 - \lambda_2)x = 0.$$

Case 5. Suppose $[\![\gamma_i]\!](x) = \sqrt{\frac{2}{\pi}}\frac{1}{a_i^3}x^2 \exp -x^2/(2a_i^2)$ for $i = 1, 2$. We have

$$\lim_{x\to\infty} \frac{\sqrt{\frac{2}{\pi}}\frac{1}{a_i^3}x^2 \exp -x^2/(2a_i^2)}{x^2 \exp -x^2/(2a_i^2)} = 0 \tag{5.4}$$

for $i = 1, 2$. Assume $a_1 < a_2$ then

$$\lim_{x\to\infty} [\![\gamma_1]\!](x)/[\![\gamma_2]\!](x) = \lim_{x\to\infty} \frac{\exp -x^2/(2a_1^2)}{\exp -x^2/(2a_2^2)} = \lim_{x\to\infty} \exp \frac{1}{2}(a_2^{-2} - a_1^{-2})x^2 = 0.$$

Case 6. Suppose $[\![\gamma_i]\!](x) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp -\frac{(x-\mu_i)^2}{2\sigma_i^2}$ for $i = 1, 2$. We have

$$\lim_{x\to\infty} \frac{\frac{1}{\sigma_i\sqrt{2\pi}} \exp -\frac{(x-\mu_i)^2}{2\sigma_i^2}}{\exp -\frac{(x-\mu_i)^2}{2\sigma_i^2}} = 1 \tag{5.5}$$

as $x \to \infty$ for $i = 1, 2$. Further, we have

$$\begin{aligned}\frac{\exp -\frac{(x-\mu_1)^2}{2\sigma_1^2}}{\exp -\frac{(x-\mu_2)^2}{2\sigma_2^2}} &= \exp\Big(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\Big) \\ &= \exp\Big(\big(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\big)x^2 - \big(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\big)x + \big(\frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}\big)\Big). \end{aligned} \tag{5.6}$$

We consider the coefficients of the polynomial in (5.6). If $\sigma_1 < \sigma_2$ then $\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2} < 0$ and so $[\![\gamma_1]\!] \prec [\![\gamma_2]\!]$. If $\sigma_1 = \sigma_2$ but $\mu_1 < \mu_2$ then $\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2} > 0$ and so $[\![\gamma_1]\!] \prec [\![\gamma_2]\!]$. These cases are enough to show that either $[\![\gamma_1]\!] \prec [\![\gamma_2]\!]$ or $[\![\gamma_2]\!] \prec [\![\gamma_1]\!]$ if $\sigma_1 \neq \sigma_2$ or $\mu_1 \neq \mu_2$.

Next, we consider some more cases:

7. $\gamma_1 \in \Gamma_m$ and $\gamma_2 \in \Gamma_t$,
8. $\gamma_1 \in \Gamma_t$ and $\gamma_2 \in \Gamma_\chi$,
9. $\gamma_1 \in \Gamma_\chi$ and $\gamma_2 \in \Gamma_e$,
10. $\gamma_1 \in \Gamma_e$ and $\gamma_2 \in \Gamma_{mb}$,
11. $\gamma_1 \in \Gamma_{mb}$ and $\gamma_2 \in \Gamma_g$.

In each case we show that $[\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) \to 0$ as $x \to \infty$ and hence $[\![\gamma_2]\!] \prec [\![\gamma_1]\!]$.

Case 7. Suppose $[\![\gamma_1]\!](x) = x^k$ and $[\![\gamma_2]\!](x) = \frac{\tilde{\Gamma}((\nu+1)/2)}{\sqrt{\nu\pi}\tilde{\Gamma}(\nu/2)}(1 + \frac{x^2}{\nu})^{-(\nu+1)/2}$ where $\nu \in \mathbb{Q}_+$ and $k \in \mathbb{N}_0$. Then by (5.3),

$$\lim_{x\to\infty} [\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) = \lim_{x\to\infty} \frac{x^{-(\nu+1)}}{x^k} = \lim_{x\to\infty} x^{-\nu-1-k} = 0.$$

Case 8. Suppose $[\![\gamma_1]\!](x) = \frac{\tilde{\Gamma}((\nu+1)/2)}{\sqrt{\nu\pi}\tilde{\Gamma}(\nu/2)}(1+\frac{x^2}{\nu})^{-(\nu+1)/2}$ and $[\![\gamma_2]\!](x) = \frac{1}{2^{k/2}\tilde{\Gamma}(k/2)}x^{k/2-1}\exp -x/2$ where $\nu \in \mathbb{Q}_+$ and $k \in \mathbb{N}$. Then by (5.3) and (5.3),

$$\lim_{x\to\infty} [\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) = \lim_{x\to\infty} \frac{x^{k/2-1}\exp -x/2}{x^{-(\nu+1)}} = \lim_{x\to\infty} x^{k/2+\nu}\exp -x/2 = 0.$$

Case 9. Suppose $[\![\gamma_1]\!](x) = \frac{1}{2^{k/2}\tilde{\Gamma}(k/2)}x^{k/2-1}\exp -x/2$ and $[\![\gamma_2]\!](x) = \lambda\exp -\lambda x$ where $k \in \mathbb{N} \setminus \{1\}$ and $\lambda \in \mathbb{Q}_+$. In the case $\lambda = \frac{1}{2}$ and $k = 2$,

$$[\![\gamma_1]\!](x) = \frac{1}{2\tilde{\Gamma}(1)}\exp -x/2 = \frac{1}{2}\exp -x/2 = [\![\gamma_2]\!](x).$$

Therefore, $\lambda(\{x \in \mathbb{R} \mid [\![\gamma_1]\!](x) \neq [\![\gamma_2]\!](x)\}) > 0$ implies either $\lambda \neq \frac{1}{2}$ or $k \neq 2$. By (5.3),

$$\lim_{x\to\infty} [\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) = \lim_{x\to\infty} \frac{\lambda\exp -\lambda x}{x^{k/2-1}\exp -x/2} = \lim_{x\to\infty} \lambda x^{-(k/2-1)}\exp -(\lambda - \frac{1}{2})x.$$

In the case $\lambda < \frac{1}{2}$ then $[\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) \to \infty$ as $x \to \infty$. In the case $\lambda > \frac{1}{2}$ then $[\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) \to 0$ as $x \to \infty$. In the case $\lambda = \frac{1}{2}$ and $k \in \{0, 1\}$ we have $[\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) \to \infty$ as $x \to \infty$. In the case $\lambda = \frac{1}{2}$ and $k > 2$ we have $[\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) \to 0$ as $x \to \infty$. These cases are enough to establish that either $[\![\gamma_1]\!] \prec [\![\gamma_2]\!]$ or $[\![\gamma_2]\!] \prec [\![\gamma_1]\!]$.

Case 10. Suppose $[\![\gamma_1]\!](x) = \lambda\exp -\lambda x$ and $[\![\gamma_2]\!](x) = \sqrt{\frac{2}{\pi}}\frac{1}{a^3}x^2\exp -x^2/(2a^2)$ where $\lambda, a \in \mathbb{Q}_+$. Then by (5.4),

$$\lim_{x\to\infty} [\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) = \lim_{x\to\infty} \frac{x^2\exp -x^2/(2a^2)}{\lambda\exp -\lambda x} = \lim_{x\to\infty} \frac{1}{\lambda}\exp\Big(-\frac{x^2}{2a^2} + \lambda x\Big) = 0.$$

Case 11. Suppose $[\![\gamma_1]\!](x) = \sqrt{\frac{2}{\pi}}\frac{1}{a^3}x^2\exp -x^2/(2a^2)$ and $[\![\gamma_2]\!](x) = \frac{1}{\sigma\sqrt{2\pi}}\exp -\frac{(x-\mu)^2}{2\sigma^2}$ where $a, \sigma \in \mathbb{Q}_+$ and $\mu \in \mathbb{Q}$. Then by (5.5) and (5.4),

$$\lim_{x\to\infty} [\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) = \lim_{x\to\infty} \frac{\exp -\frac{(x-\mu)^2}{2\sigma^2}}{x^2\exp -x^2/(2a^2)} = \lim_{x\to\infty} \frac{1}{x^2}\exp\Big(\frac{x^2}{2a^2} - \frac{(x-\mu)^2}{2\sigma^2}\Big)$$

Since $\frac{x^2}{2a^2} - \frac{(x-\mu)^2}{2\sigma^2}$ is a polynomial with degree 2 or less we have $\lim_{x\to\infty} [\![\gamma_2]\!](x)/[\![\gamma_1]\!](x) \in \{0, \infty\}$ and thus either $[\![\gamma_1]\!] \prec [\![\gamma_2]\!]$ or $[\![\gamma_2]\!] \prec [\![\gamma_1]\!]$.

Cases 1-11 together with the transitivity property of $\prec$ imply that $\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g$ is asymptotically ordered. □

Proposition 5.1.9. Γ_{GEM} *is linearly decomposable and joint graph decomposable.*

Proof of Proposition 5.1.9. Let γ be a profile in Γ_{GEM}. Then,

$$[\![\gamma]\!] = \sum_{i=1}^{I} r_i m_i + \sum_{j=1}^{J} s_j g_j + \sum_{k=1}^{K} t_k e_k$$

where $I, J, K \in \mathbb{N}$, $r_i, s_j, t_k \in \mathbb{Q}$ for $i \in [I]$, $j \in [J]$ and $k \in [K]$ respectively. Further, m_i, g_j, e_k are distinct pdfs of interval-domain monomial, gaussian and exponential distributions for $i \in [I]$, $j \in [J]$ and $k \in [K]$ respectively. Let $m_i(x) = x^{n_i}\chi_{[a_i,b_i)}(x)$, $g_j(x) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp -\frac{(x-\mu_j)^2}{2\sigma_j^2}$ and $e_k(x) = \lambda_k \chi_{(0,\infty)}(x) \exp(-\lambda_k x)$ for all $x \in \mathbb{R}$ and $i \in [I]$, $j \in [J]$ and $k \in [K]$ respectively. Let $\alpha_i \in \Gamma_m$ encode $[\![\alpha_i]\!](x) = x^{n_i}$ for $i \in [I]$. Let $\beta_j \in \Gamma_g$ encode $[\![\beta_j]\!](x) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp -\frac{(x-\mu_j)^2}{2\sigma_j^2}$ for $j \in [J]$. Let $\gamma_k \in \Gamma_e$ encode $[\![\gamma_k]\!](x) = \lambda_k \exp(-\lambda_k x)$ for $k \in [K]$. We have $[\![\mathcal{I}(\alpha_i, a_i, b_i)]\!] = m_i$, $[\![\beta_j]\!] = g_j$ and $[\![\mathcal{I}(\gamma_k, 0, \infty)]\!] = e_k$ up to a λ_{Leb}-null set for $i \in [I]$, $j \in [J]$ and $k \in [K]$ respectively. Let

$$\begin{aligned} \gamma' = \mathcal{S}(&r_1, \mathcal{I}(\alpha_1, a_1, b_1), \ldots, r_I, \mathcal{I}(\alpha_I, a_I, b_I), \\ &s_1, \beta_1, \ldots, s_J, \beta_J, \\ &t_1, \mathcal{I}(\gamma_1, 0, \infty), \ldots, t_K, \mathcal{I}(\gamma_K, 0, \infty)) \end{aligned} \tag{5.7}$$

and then $[\![\gamma]\!] = [\![\gamma']\!]$ up to a λ_{Leb}-null set. The algorithm that produces γ' from γ takes polynomial time in the size of the encoding of γ.

We now show Γ_{GEM} is linearly decomposable. Consider the finite subset $\Gamma_0 \subseteq \Gamma_{GEM}$. In polynomial time we may compute an equivalent finite subset $\Gamma_0' \in$ Span-$\mathcal{I}(\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g)$ using the construction (5.7). Since Span-$\mathcal{I}(\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g)$ is linearly decomposable we may perform the linear decomposition algorithm on Γ_0' and then map the profiles obtained in the result back to the corresponding profiles in Γ_0.

Finally we show Γ_{GEM} is joint graph decomposable. Let $\gamma_1, \gamma_2 \in \Gamma_{GEM}$ and let γ_1', γ_2' be obtained by the construction (5.7). We have

$$\lambda_{Leb}(\mathrm{supp}\,([\![\gamma_1]\!]) \cap \mathrm{supp}\,([\![\gamma_2]\!])) = \lambda_{Leb}(\mathrm{supp}\,([\![\gamma_1']\!]) \cap \mathrm{supp}\,([\![\gamma_2']\!]))$$

and so Γ_{GEM} is joint graph decomposable because Span-$\mathcal{I}(\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g)$ is joint graph decomposable. $\square$

5.4 Dirichlet Distributions

In this section we define a profile language Γ_D^k for each integer $k \geq 2$ which can encode Dirichlet distributions. Up to this point we have only considered sets of observations with $\Sigma \subseteq \mathbb{R}$ but in this section we consider observations from the $k-1$ simplex, Δ^k. We will show that Γ_D^k is linearly decomposable and joint graph decomposable which demonstrates that our results about equivalence and distinguishability can apply when the set of observations is multi-dimensional. In particular we show that $[\![\Gamma_D^k]\!]$ is a linearly independent set of functions. Towards this goal we parametrise the set of observations, writing $\vec{x}(d) \in \Delta^k$ where $d \in (0, \infty)$. We use a lexicographic ordering on the parameters of the Dirichlet distributions inspired by Case 6 in the proof of Lemma 5.3.3. By ordering the elements of an assumed linear dependence between distinct Dirichlet pdfs according to the lexicographic ordering on the parameters, we show that each element of the linear dependence has a different growth rate as $d \to \infty$. This contradicts the assumption that such a linear dependence exists.

Example 5.4.1. Let $k \in \mathbb{N}$ with $k \geq 2$. The $k-1$ simplex is $\Delta^k := \{(x_1, \ldots, x_k) \in [0,1]^k \mid \sum_{i=1}^k x_i = 1\}$ and a Dirichlet distribution of order k with parameters $\alpha_1 \cdots \alpha_k \in \mathbb{Q}_+$ has probability density function $f : \Delta^k \to (0, \infty)$ where

$$f(x_1, \ldots, x_k) = \frac{1}{B(\alpha_1, \cdots, \alpha_k)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}.$$

The pdf f is with respect to the Lebesgue measure λ_{Leb}^{k-1} on $[0,1]^{k-1}$. Here, $B : \mathbb{R}^k \to \mathbb{R}$ is the multivariate beta function:

$$B(\alpha_1, \ldots, \alpha_k) = \frac{\prod_{i=1}^k \tilde{\Gamma}(\alpha_i)}{\tilde{\Gamma}\left(\sum_{i=1}^k \alpha_i\right)}$$

where $\tilde{\Gamma}$ is the gamma function defined in Section 5.1.

For fixed $k \geq 2$, we define a constructor $D_k(\alpha_1, \ldots, \alpha_k)$ for encoding a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$. We define the profile language

$$\Gamma_D^k = \{D_k(\alpha_1, \ldots, \alpha_k) \mid \alpha_1, \ldots, \alpha_k \in \mathbb{Q}_+\}.$$

Let $n \in \mathbb{N}$ for each $j \in [n]$ let $\alpha_1^j, \ldots, \alpha_k^j \in \mathbb{Q}_+$ such that for each $j, j' \in [n]$ there exists $i \in [k]$ such that $\alpha_i^j \neq \alpha_i^{j'}$. We will show that for profiles $D_k(\alpha_1^1, \ldots, \alpha_k^1), \ldots, D_k(\alpha_1^n, \ldots, \alpha_k^n) \in \Gamma_D^k$ the set

$$\{[\![D_k(\alpha_1^1, \ldots, \alpha_k^1)]\!], \ldots, [\![D_k(\alpha_1^n, \ldots, \alpha_k^n)]\!]\}$$

is linearly independent.

We define a lexicographic order on $D_k(\alpha_1^1, \ldots, \alpha_k^1), \ldots, D_k(\alpha_1^n, \ldots, \alpha_k^n)$. For $j, j' \in [n]$ we say that $D_k(\alpha_1^j, \ldots, \alpha_k^j) <_{\text{lex}} D_k(\alpha_1^{j'}, \ldots, \alpha_k^{j'})$ if

$$\left(\alpha_1^j < \alpha_1^{j'}\right) \vee \bigvee_{i=2}^{k} \left(\left(\alpha_i^j < \alpha_i^{j'}\right) \wedge \bigwedge_{r=1}^{i-1} \left(\alpha_r^j = \alpha_r^{j'}\right)\right).$$

Without loss of generality assume $D_k(\alpha_1^1, \ldots, \alpha_k^1) <_{\text{lex}} \cdots <_{\text{lex}} D_k(\alpha_1^n, \ldots, \alpha_k^n)$. Let $1 \leq j < j' \leq k$ so that $D_k(\alpha_1^j, \ldots, \alpha_k^j) <_{\text{lex}} D_k(\alpha_1^{j'}, \ldots, \alpha_k^{j'})$ and there is $i_0 \in [k]$ such that $\alpha_{i_0}^j < \alpha_{i_0}^{j'}$ and if $i_0 > 1$ then $\alpha_i^j = \alpha_i^{j'}$ for all $i < i_0$. We have that

$$\lim_{d \to \infty} \sum_{i=1}^{k} d^{k+1-i}(\alpha_i^j - \alpha_i^{j'}) = \lim_{d \to \infty} \sum_{i=i_0}^{k} d^{k+1-i}(\alpha_i^j - \alpha_i^{j'}) = -\infty \tag{5.8}$$

since $\alpha_{i_0}^j - \alpha_{i_0}^{j'} < 0$.

Assume there are coefficients $s_1, \ldots, s_n \in \mathbb{R} \setminus \{0\}$ such that

$$\sum_{j=1}^{n} s_j [\![D_k(\alpha_1^j, \ldots, \alpha_k^j)]\!] = 0.^1$$

It follows that

$$\sum_{j=1}^{n-1} \left(s_j [\![D_k(\alpha_1^j, \ldots, \alpha_k^j)]\!] / [\![D_k(\alpha_1^n, \ldots, \alpha_k^n)]\!]\right) = -s_n.$$

For $d \in (0, \infty)$ let $t(d) = \sum_{i=1}^{k} e^{d^i}$ and consider the parametrised row vector $\vec{x}(d) := (e^{d^k}, \ldots, e^{d^2}, e^{d^1})/t(d) \in \Delta^k$. For $1 \leq j < n$ we have

$$\begin{aligned}
\lim_{d \to \infty} [\![D_k(\alpha_1^j, \ldots, \alpha_k^j)]\!](\vec{x}(d)) / [\![D_k(\alpha_1^1, \ldots, \alpha_k^1)]\!](\vec{x}(d)) &= \lim_{d \to \infty} \prod_{i=1}^{k} \frac{e^{d^{k+1-i}(\alpha_i^j - 1)}}{e^{d^{k+1-i}(\alpha_i^n - 1)}} \frac{t(d)}{t(d)} \\
&= \lim_{d \to \infty} \prod_{i=1}^{k} \exp\left(d^{k+1-i}(\alpha_i^j - \alpha_i^n)\right) \\
&= \lim_{d \to \infty} \exp\left(\sum_{i=1}^{k} d^{k+1-i}(\alpha_i^j - \alpha_1^n)\right) \\
&= 0 \text{ by (5.8).}
\end{aligned}$$

[1] Previously we required this linear dependence holds almost everywhere. Here, this condition is not necessary since Dirichlet pdfs are continuous.

Hence, $s_n = 0$, a contradiction. We conclude that $[\![\Gamma_D^k]\!]$ is linearly independent. It follows that Γ_D^k is linearly decomposable since we may define a trivial procedure that takes a subset $\{\gamma_1, \ldots, \gamma_n\} \subset \Gamma_D^k$ and returns profiles $\beta_i := \gamma_i$ for all $i \in [n]$ and coefficients $b_{i,j} := \delta_{i,j}$ where $\delta_{i,j}$ is the Kronecker delta function.

Also, Γ_D^k is joint graph decomposable since we may define a trivial procedure that takes two profiles $\gamma_1, \gamma_2 \in \Gamma_D^k$ and always returns true. This is because $[\![\gamma_1]\!](\vec{x}) > 0$ and $[\![\gamma_2]\!](\vec{x}) > 0$ for all $\vec{x} \in (0,1)^{k-1} \cap \Delta^k$ and $\lambda_{Leb}^{k-1}(\Delta^k) = \lambda_{Leb}^{k-1}((0,1)^{k-1})$.

6 The Sequential Probability Ratio Test

In previous chapters we introduced the total variation distance as a metric for capturing probabilistic similarity of a HMM started from two different initial distributions. We defined what it means for two initial distributions to be distinguishable. We show in this chapter that for distinguishable initial distributions one can use the sequential probability ratio test (SPRT) [27] to decide which initial distribution a particular run belongs to. The execution time of this algorithm is itself a random variable that depends on the run. Moreover, the distribution of the execution time depends on the specific HMM and also the initial distributions. Even though two initial distributions are distinguishable, they may be distinguishable to different degrees as observed by the execution time of the SPRT.

The word "ratio" in the SPRT refers to the *likelihood ratio* of two initial distributions. The likelihood ratio converges to 0 at a particular rate. We show in Chapter 7 that there are quadratically many such rates (in the number of states): each of which realised with a certain probability. We call these rates *likelihood exponents*.

For distinguishable initial distributions we show that the likelihood exponent and the execution time of the SPRT are, broadly speaking, inversely proportional. Regarding the execution time of the SPRT, we examine its expectation in Section 6.1 and then its distribution in Section 6.3. We give computational results for the distribution of the likelihood exponents in Section 6.4. In Section 6.5 we show how the distribution of the likelihood exponent correspond to the hitting probabilities of different bottom SCCs in a Markov chain we refer to as the *Numerator Support Markov Chain*. In Section 6.6 we give proofs of the various ways that the likelihood exponents and execution time of the SPRT are linked.

In both this chapter and also chapter 7 our results only apply to finite-observation chains since as far as we are aware, these results are novel even in this case. From now on, we assume that a HMM has finite-observations and hence will be given the label $\mathcal{F}$.

Let $\mathcal{F} = (Q, \Sigma, M)$ be a HMM and let π_1 and π_2 be initial distributions. For $n \in \mathbb{N}$, the *likelihood ratio* L_n is a random variable on Σ^ω given by $L_n(w) = \frac{\|\pi_1 M(w_n)\|}{\|\pi_2 M(w_n)\|}$. Based on results from [23] we have the following lemma.

Lemma 6.0.1. *Let π_1, π_2 be initial distributions.*

1. $\lim_{n\to\infty} L_n$ *exists* $\mathbb{P}_{\pi_2}$*-almost surely and lies in* $[0, \infty)$.
2. $\lim_{n\to\infty} L_n = 0$ $\mathbb{P}_{\pi_2}$*-almost surely if and only if* π_1 *and* π_2 *are distinguishable.*

Proof. The first part is [23, Proposition 6]. Towards the second part, the following equalities hold.

$$\begin{aligned}
1 - d(\pi_1, \pi_2) &= \lim_{n\to\infty} \sum_{w\in\Sigma^n} \min\{\|\pi_1 M(w)\|, \|\pi_2 M(w)\|\} && \text{by [23, Theorem 7]} \\
&= \lim_{n\to\infty} \sum_{w\in\Sigma^n} \min\{L_n(w), 1\} \|\pi_2 M(w)\| \\
&= \lim_{n\to\infty} \mathbb{E}_{\pi_2}\Big[\min\{L_n, 1\}\Big] \\
&= \mathbb{E}_{\pi_2}\Big[\lim_{n\to\infty} \min\{L_n, 1\}\Big] && \text{as } 0 \le \min\{L_n(w), 1\} \le 1.
\end{aligned}$$

Then, $\lim_{n\to\infty} \min\{L_n, 1\} = 0 \iff \lim_{n\to\infty} L_n = 0$. □

Example 6.0.2. We illustrate convergence of the likelihood ratio using an example from [37] where the authors use HMMs to model sleep cycles. They took measurements of 51 healthy and 51 diseased individuals and using electrodes attached to the scalp, they read electrical signal data as part of an electroencephalography (EEG) during sleep. They split the signal into 30 second intervals and mapped each interval onto the simplex $\Delta^3 = \{(x_1, x_2, x_3, x_4) \in [0, 1]^4 \mid \sum_{i=1}^4 x_i = 1\}$. For each individual this results in a time series of points in Δ^3. They modelled this data using two HMMs, each with 5 states, for healthy and diseased individuals using a numerical maximum likelihood estimate. Each state is associated with a probability density function describing the distribution of observations in Δ^3.

The two embedded Markov chains have the following transition matrices:

$$T_1 = \begin{bmatrix} 0.793 & 0.099 & 0.035 & 0.064 & 0.009 \\ 0.078 & 0.769 & 0.006 & 0.144 & 0.003 \\ 0.018 & 0.004 & 0.833 & 0.134 & 0.012 \\ 0.022 & 0.094 & 0.054 & 0.827 & 0.002 \\ 0.011 & 0.005 & 0.035 & 0.005 & 0.945 \end{bmatrix}, T_2 = \begin{bmatrix} 0.641 & 0.109 & 0.031 & 0.040 & 0.015 \\ 0.202 & 0.699 & 0.008 & 0.089 & 0.003 \\ 0.026 & 0.002 & 0.823 & 0.062 & 0.035 \\ 0.123 & 0.189 & 0.114 & 0.808 & 0.016 \\ 0.007 & 0.001 & 0.024 & 0.001 & 0.931 \end{bmatrix}.$$

Their HMMs are state-labelled. For each state i, they fit a Dirichlet pdf f_i describing the distribution of observations in Δ^3 emitted at state i. The pdfs of diseased and healthy individuals were so similar that they used the same pdf for both HMMs. Thus the two HMMs differ only in the transition probabilities.

Since Δ^3 is infinite and in this chapter we assume finite observation alphabets, we partition the simplex into the sets

$$U_k = \{x \in \Delta^3 \mid f_k(x) \geq \sup_i f_i(x)\}$$

for $k = 1, \ldots, 5$. The set U_k contains the points in Δ^3 most likely to be produced in state k. We assign a letter a_k for each U_k, and define a set of observations $\Sigma = \{a_1, \ldots, a_5\}$. Thus, the probability of producing letter a_k from state i is given as $O_{i,k} = \int_{U_k} f_i(x)\, dx$. We estimated the entries of O using a numerical Monte Carlo technique. We generated 100,000 samples from all 5 Dirichlet distributions in their paper which yielded the estimate

$$O = \begin{pmatrix} 0.9172 & 0.0803 & 0 & 0.0002 & 0.0024 \\ 0.0719 & 0.8606 & 0 & 0.0665 & 0.0010 \\ 0 & 0.0007 & 0.8546 & 0.1055 & 0.0392 \\ 0.0008 & 0.0998 & 0.0663 & 0.8257 & 0.0075 \\ 0.0109 & 0.0094 & 0.1046 & 0.0334 & 0.8416 \end{pmatrix}.$$

Since we consider transition labelled HMMs, we define transition functions M_1, M_2 with

$$M_m(a_k)_{i,j} = \left(T_m\right)_{i,j} O_{i,k}$$

for $m = 1, 2$. Let $Q = [10]$. We construct the HMM $\mathcal{F} = (Q, \Sigma, M)$ where

$$M(a) = \begin{pmatrix} M_1(a) & 0 \\ 0 & M_2(a) \end{pmatrix}$$

for each $a \in \Sigma$.

Let π_1 and π_2 be the Dirac distributions on states 1 and 6 respectively. These initial distributions correspond to healthy and diseased individuals started from sleep state 1.

Using the algorithm from [23] one can show that π_1 and π_2 are distinguishable. We sampled runs of $\mathcal{F}$ started from π_1 and π_2 and plotted the corresponding sequences of $\ln L_n$. We refer to each of these two plots as a *log-likelihood plot*; see Figure 6.1.

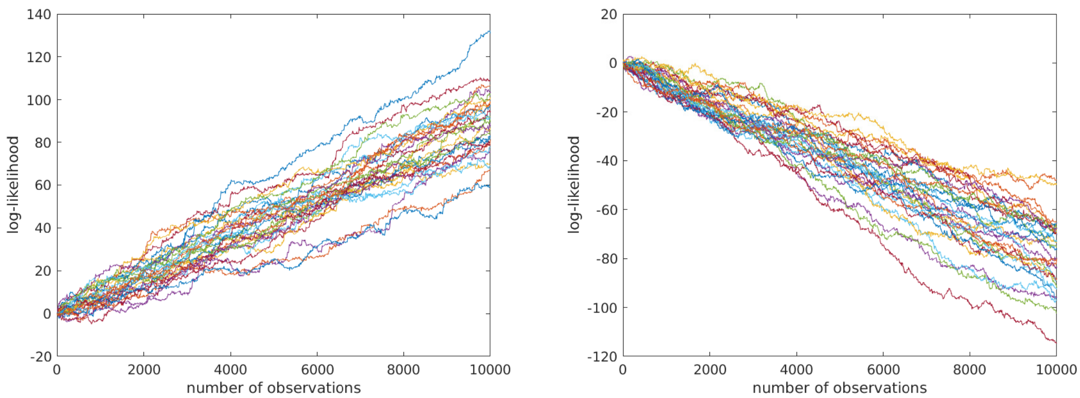


Figure 6.1: The two images show two log-likelihood plots of sample runs produced by π_1 and π_2, respectively.

By the second point of Lemma 6.0.1 it follows that $\ln L_n$ converges $\mathbb{P}_{\pi_1}$-a.s. (almost-surely) to ∞ and $\mathbb{P}_{\pi_2}$-a.s. to $-\infty$. This is affirmed by Figure 6.1. Both log-likelihood plots also appear to follow a particular slope. This suggests that we can distinguish between words produced by π_1 and π_2 by tracking the value of $\ln L_n$ to see whether it crosses a lower or upper threshold. This is the intuition behind the *Sequential Probability Ratio Test* (SPRT).

6.1 Sequential Probability Ratio Test

Fix a HMM $\mathcal{F} = (Q, \Sigma, M)$ for the rest of the thesis. Given initial distributions π_1, π_2 and error bounds $\alpha, \beta \in (0,1)$, the SPRT runs as follows. It continues to read observations and computes the value of $\ln L_n$ until $\ln L_n$ leaves the interval $[A, B]$, where $A := \ln \frac{\alpha}{1-\beta}$ and $B := \ln \frac{1-\alpha}{\beta}$. If $\ln L_n \leq A$ the test outputs "π_2" and if $\ln L_n \geq B$ the test outputs "π_1". We may view the SPRT as a random variable

$\mathrm{SPRT}_{\alpha,\beta} : \Sigma^\omega \to \{\pi_1, \pi_2, ?\}$, where ? denotes that the SPRT does not terminate, i.e., $\ln L_n \in [A, B]$ for all n. We have the following property.

Proposition 6.1.1. *Suppose π_1 and π_2 are distinguishable. Let $\alpha := \mathbb{P}_{\pi_1}(\mathrm{SPRT}_{\alpha,\beta} = \pi_2)$ and $\beta := \mathbb{P}_{\pi_2}(\mathrm{SPRT}_{\alpha,\beta} = \pi_1)$. Then for error bounds $A, B \in \mathbb{R}$ with $A < 0 < B$ we have $A \geq \ln \frac{\alpha}{1-\beta}$ and $B \leq \ln \frac{1-\alpha}{\beta}$.*

Define the stopping time

$$N_{\alpha,\beta} \ := \ \min\{n \in \mathbb{N} \mid \ln L_n \notin [A, B]\} \ \in \ \mathbb{N} \cup \{\infty\}\,.$$

We have that $N_{\alpha,\beta}$ is monotone decreasing in the sense that for $\alpha \leq \alpha'$ and $\beta \leq \beta'$ we have $N_{\alpha,\beta} \geq N_{\alpha',\beta'}$. When π_1 and π_2 are distinguishable, $N_{\alpha,\beta}$ is $\mathbb{P}_{\pi_2}$-a.s. finite by the second point of Lemma 6.0.1.

Proof of Proposition 6.1.1. Let $a = \frac{\alpha}{1-\beta}$ and $b = \frac{1-\alpha}{\beta}$. Write $N := N_{\alpha,\beta}$ and let $W_n^1 = \{w \in \Sigma^\omega \mid a \leq L_m(w) \leq b \ \forall m < n, L_n < a\}$ then

$$\begin{aligned}\mathbb{P}_{\pi_1}\Big(L_N < a\Big) &= \sum_{n=1}^{\infty} \mathbb{P}_{\pi_1}\Big(W_n^1\Big) = \sum_{n=1}^{\infty} \sum_{w \in W_n^1} \pi_1 M(w) \overline{1}^\top = \sum_{n=1}^{\infty} \sum_{w \in W_n^1} L_n(w) \pi_2 M(w) \overline{1}^\top \\ &\leq a \sum_{n=1}^{\infty} \sum_{w \in W_n^1} \pi_2 M(w) \overline{1}^\top = a \sum_{n=1}^{\infty} \mathbb{P}_{\pi_2}\Big(W_n^1\Big) = a \mathbb{P}_{\pi_2}\Big(L_N < a\Big).\end{aligned}$$

Similarly, we may derive $\mathbb{P}_{\pi_2}\Big(L_N > b\Big) \leq \frac{1}{b} \mathbb{P}_{\pi_1}\Big(L_N > b\Big)$ so that

$$\begin{aligned} a &\geq \frac{\mathbb{P}_{\pi_1}\Big(L_N < a\Big)}{\mathbb{P}_{\pi_2}\Big(L_N < a\Big)} = \frac{\mathbb{P}_{\pi_1}\Big(L_N < a\Big)}{1 - \mathbb{P}_{\pi_2}\Big(L_N > b\Big)} \\ b &\leq \frac{\mathbb{P}_{\pi_1}\Big(L_N > b\Big)}{\mathbb{P}_{\pi_2}\Big(L_N > b\Big)} = \frac{1 - \mathbb{P}_{\pi_1}\Big(L_N < a\Big)}{\mathbb{P}_{\pi_2}\Big(L_N > b\Big)} \end{aligned}$$

and the result follows. □

In the following we consider the SPRT with respect to the measure $\mathbb{P}_{\pi_2}$. This is without loss of generality as there is a dual version of the SPRT, say $\overline{\mathrm{SPRT}}$ with $\overline{L}_n = 1/L_n$ instead of L_n, such that $\overline{\mathrm{SPRT}}_{\beta,\alpha} = \mathrm{SPRT}_{\alpha,\beta}$.

6.2 Expectation of $N_{\alpha,\beta}$

Consider the two-state HMM where $p_1 \neq p_2$.

$$\begin{tikzcd} s_1 \arrow[loop left, "p_1 : a"] \arrow[loop right, "(1-p_1) : b"] & & & s_2 \arrow[loop left, "p_2 : a"] \arrow[loop right, "(1-p_2) : b"] \end{tikzcd}$$

(The Dirac distributions of) s_1 and s_2 are distinguishable. Further, the increments $\ln L_{n+1} - \ln L_n$ are independent and identically distributed (i.i.d.) and $0 > \mathbb{E}_{s_2}[\ln L_{n+1} - \ln L_n] = p_2 \ln \frac{p_1}{p_2} + (1-p_2) \ln \frac{1-p_1}{1-p_2} =: \ell$. Intuitively as ℓ gets more negative, the HMMs become more different.[1] Indeed, Wald [27] shows that the expected stopping time $\mathbb{E}_{s_2}[N_{\alpha,\beta}]$ and ℓ are inversely proportional:

$$\mathbb{E}_{s_2}[N_{\alpha,\beta}] = \frac{\beta \ln \frac{1-\alpha}{\beta} + (1-\beta) \ln \frac{\alpha}{1-\beta}}{\ell}. \tag{6.1}$$

This Wald formula cannot hold in general for (multi-state) HMMs. The increments $\ln L_{n+1} - \ln L_n$ need not be independent and $\mathbb{E}_{s_2}[\ln L_{n+1} - \ln L_n]$ can be different for different n. Further, $|\ln L_{n+1} - \ln L_n|$ can be unbounded; cf. [24, Example 6].

Nevertheless, in Figure 6.1 we observed that $\ln L_n$ appears to decrease linearly (on the π_2 plot). Indeed, we show in Theorem 6.2.3 below that the limit $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ exists $\mathbb{P}_{\pi_2}$-almost surely. Intuitively it corresponds to the average slope of the log-likelihood plot for π_2. In the two-state case, there is a simple proof of this using the law of large numbers:

$$\lim_{n\to\infty} \frac{1}{n} \ln L_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} [\ln L_{i+1} - \ln L_i] = \mathbb{E}_{\pi_2}[\ln L_1 - \ln L_0] = \ell \quad \mathbb{P}_{\pi_2}\text{-a.s.}$$

The number ℓ is called a likelihood exponent, as defined generally in the following definition.

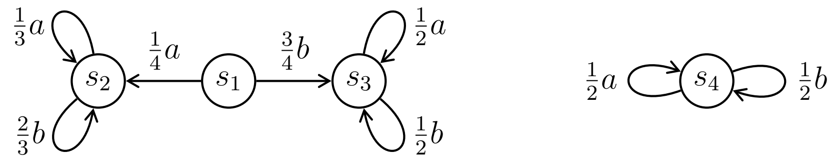
Definition 6.2.1. For initial distributions π_1, π_2, a number $\ell \in [-\infty, 0]$ is a *likelihood exponent* if $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n = \ell) > 0$.

By the first point of lemma 6.0.1 we have $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n > 0) = 0$, as $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} L_n < \infty) = 1$. Hence, we may restrict likelihood exponents to $[-\infty, 0]$. We write $\Lambda_{\pi_1,\pi_2} \subseteq [-\infty, 0]$ for the set of likelihood exponents for π_1, π_2 and define $\Lambda := \bigcup_{\pi_1,\pi_2} \Lambda_{\pi_1,\pi_2}$; i.e., Λ depends only on the HMM $\mathcal{F}$. For $\ell \in \Lambda$ we define the event $E_\ell = \{\lim_{n\to\infty} \frac{1}{n} \ln L_n = \ell\}$.

[1] In fact, ℓ is the *KL-divergence* of the distributions f_1, f_2 where $f_i(a) = p_i$ and $f_i(b) = 1 - p_i$ for $i = 1, 2$.

Example 6.2.1. In the case of Example 6.0.2 we have $\Lambda_{\pi_1,\pi_2} = \{\ell\}$ where the slope of the right hand side of Figure 6.1 suggests that $\ell \approx -\frac{80}{10000} = -0.008$.

Example 6.2.2. Even for fixed π_1, π_2 there may be multiple likelihood exponents. Consider the following HMM with initial Dirac distributions $\pi_1 = e_{s_1}$ and $\pi_2 = e_{s_4}$.



We observe two different likelihood exponents depending on the first letter produced. If the first letter is a then $\ln L_{n+1} - \ln L_n$ are i.i.d. for $n \geq 1$ and $\lim_{n\to\infty} \frac{1}{n} \ln L_n = \frac{1}{2} \ln \frac{1/3}{1/2} + \frac{1}{2} \ln \frac{2/3}{1/2} = \frac{1}{2} \ln \frac{8}{9} =: \ell$ like the two-state example above with $p_1 = \frac{1}{3}$ and $p_2 = \frac{1}{2}$. If the first letter is b then $L_n = \frac{3}{2}$ for all $n \geq 1$ and $\lim_{n\to\infty} \frac{1}{n} \ln L_n = 0$. Thus, $\Lambda_{\pi_1,\pi_2} = \{\ell, 0\}$ and $\mathbb{P}_{\pi_2}(E_\ell) = \mathbb{P}_{\pi_2}(E_0) = \frac{1}{2}$.

The following theorem is perhaps the most fundamental contribution of this chapter. It follows from a stronger theorem, theorem 7.2.4, which we prove in chapter 7.

Theorem 6.2.3. *For any initial distributions* π_1, π_2 *the limit* $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ *exists* $\mathbb{P}_{\pi_2}$*-almost surely. Furthermore, we have* $|\Lambda| \leq |Q|^2 + 1$.

Returning to the SPRT, we investigate how $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ influences the performance of the SPRT for small α and β. Intuitively we expect a steeper slope in the likelihood plot (cf. Figure 6.1) to lead to faster termination. In the two-state case, Wald's formula (6.1) becomes

$$\mathbb{E}_{s_2}[N_{\alpha,\beta}] = \frac{\beta \ln \frac{1-\alpha}{\beta} + (1-\beta) \ln \frac{\alpha}{1-\beta}}{\ell} \sim \frac{\ln \alpha}{\ell} \text{ (as } \alpha, \beta \to 0), \tag{6.2}$$

where we use the notation $\sim$ defined as follows. For functions $f, g : (0,\infty) \times (0,\infty) \to (0,\infty)$ we write "$f(x,y) \sim g(x,y)$ (as $x, y \to 0$)" to denote that for all $\varepsilon > 0$ there is $\delta > 0$ such that for all $x, y \in (0,\delta)$ we have $f(x,y)/g(x,y) = [1-\varepsilon, 1+\varepsilon]$.

In Theorem 6.2.4 below we generalise Equation (6.2) to arbitrary HMMs. Indeed a very similar asymptotic identity holds. In the case that $\Lambda = \{\ell\}$ and $\ell \in (-\infty, 0)$ we have $\mathbb{E}_{s_2}[N_{\alpha,\beta}] \sim \frac{\ln \alpha}{\ell}$ as $\alpha, \beta \to 0$. If $|\Lambda| > 1$ then we condition our expectation on $\lim_{n\to\infty} \frac{1}{n} \ln L_n$.

Theorem 6.2.4 (Generalised Wald Formula)**.** *Let ℓ be a likelihood exponent and let π_1 and π_2 be initial distributions.*

1. *If $\ell \in (-\infty, 0)$ then $\mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] \sim \dfrac{\ln \alpha}{\ell}$ (as $\alpha, \beta \to 0$).*
2. *If $\ell = 0$ then there exist $\alpha, \beta > 0$ such that $\mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] = \infty$.*
3. *If $\ell = -\infty$ then $\sup\limits_{\alpha,\beta} \mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] < \infty$.*

The theorem above, proven in Section 6.6, pertains to the expectation of $N_{\alpha,\beta}$. In the next subsection we give additional information about the distribution of $N_{\alpha,\beta}$, further strengthening the connection between $N_{\alpha,\beta}$ and likelihood exponents.

6.3 Distribution of $N_{\alpha,\beta}$

6.3.1 Likelihood Exponent 0

Example 6.3.1. We continue with Example 6.2.2 to illustrate the second case in Theorem 6.2.4. By picking $\alpha = \frac{1}{4}, \beta = \frac{1}{4}$ the thresholds for the SPRT are $A = \ln \frac{1}{3}$ and $B = \ln 3$. If the first letter is b, then $\ln L_n = \ln \frac{3}{2}$ for all $n > 1$, thus never crosses the SPRT bounds and $\lim_{n\to\infty} \frac{1}{n} \ln L_n = 0$. Hence with probability $\frac{1}{2}$ the SPRT fails to terminate and $N_{\alpha,\beta} = \infty$. It follows that $\mathbb{P}_{\pi_2}(E_0) = \frac{1}{2}$ and $\mathbb{E}_{\pi_2}[N_{\alpha,\beta} \mid E_0] = \infty$ and, thus, $\mathbb{E}_{\pi_2}[N_{\alpha,\beta}] = \infty$.

The second part of Theorem 6.2.4 says that the expectation of $N_{\alpha,\beta}$ conditioned under E_0 is infinite. The following proposition, proven in Section 6.6, strengthens this statement. Conditioning under E_0, the probability that $N_{\alpha,\beta}$ is infinite converges to 1 as $\alpha, \beta \to 0$. Recall that $N_{\alpha,\beta}$ is monotone decreasing. It follows that $\{N_{\alpha',\beta'} = \infty\} \subseteq \{N_{\alpha,\beta} = \infty\}$ if $\alpha \leq \alpha'$ and $\beta \leq \beta'$.

Proposition 6.3.2. *The following two equalities hold up to $\mathbb{P}_{\pi_2}$-null sets:*

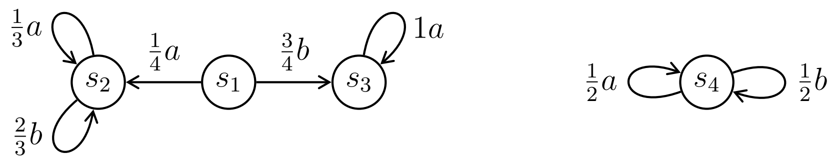
$$E_0 \;=\; \left\{\lim_{n\to\infty} L_n > 0\right\} \;=\; \bigcup_{\alpha,\beta>0} \{N_{\alpha,\beta} = \infty\}\,.$$

Thus, $\lim_{\alpha,\beta\to 0} \mathbb{P}_{\pi_2}(N_{\alpha,\beta} = \infty) = \mathbb{P}_{\pi_2}(E_0)$.

Corollary 6.3.3 (using the second point of Lemma 6.0.1)**.** *Initial distributions π_1 and π_2 are distinguishable if and only if $\mathbb{P}_{\pi_2}(E_0) = 0$ if and only if $\mathbb{P}_{\pi_2}(N_{\alpha,\beta} < \infty) = 1$ holds for all $\alpha, \beta > 0$.*

6.3.2 Likelihood Exponent $-\infty$

Example 6.3.4. Consider now a modification of Example 6.2.2 where state s_3 has the b loop removed.



The likelihood exponents are $-\infty$ and $\ell := \frac{1}{2}\ln\frac{8}{9}$ so that $\Lambda = \{-\infty, \ell\}$. Also, $\mathbb{P}_{s_4}(E_{-\infty}) = \mathbb{P}_{s_4}(E_\ell) = \frac{1}{2}$. Up to $\mathbb{P}_{s_4}$-null sets the events $E_{-\infty}$, $b\Sigma^\omega$ and $ba^*b\Sigma^\omega$ are equal. The event $ba^*b\Sigma^\omega$ represents the chain started from s_4 producing an observation which cannot be produced by the chain started from s_1, causing the SPRT to terminate for any α, β. Therefore conditioned on $E_{-\infty}$, the random variable $N_{\alpha,\beta} - 1$ is bounded by a geometric random variable with parameter $\frac{1}{2}$. Hence $\sup_{\alpha,\beta} \mathbb{E}_{\pi_2}\left[N_{\alpha,\beta} \mid E_{-\infty}\right] \le 1 + 2$.

We define the stopping time $N_\perp = \min\{n \in \mathbb{N} \mid L_n = 0\}$. Note that $\sup_{\alpha,\beta} N_{\alpha,\beta} \le N_\perp$ since $\{L_n = 0\} \subseteq \{L_n \le \frac{\alpha}{1-\beta}\}$ for all α, β. By the following proposition, the reverse inequality also holds.

Proposition 6.3.5. *The events $E_{-\infty}$ and $\{L_n = 0$ for some $n\}$ are equal. Thus, $\sup_{\alpha,\beta} N_{\alpha,\beta} = N_\perp$ and $\lim_{\alpha,\beta\to 0} \mathbb{P}_{\pi_2}(N_{\alpha,\beta} < \infty) = \mathbb{P}_{\pi_2}(E_{-\infty})$.*

Proof. The right-to-left inclusion is clear. Towards the converse, let $p_{min} > 0$ be the minimum non-zero entry in π_1 and all $M(a)$ where $a \in \Sigma$. Suppose that $L_n > 0$ holds for all n. Then we have for all $n \ge 1$:

$$\begin{aligned}\frac{1}{n}\ln L_n &= \frac{1}{n}\ln\frac{\|\pi_1 M(w_n)\|}{\|\pi_2 M(w_n)\|} \ \ge\ \frac{1}{n}\ln\|\pi_1 M(w_n)\| \ \ge\ \frac{1}{n}\ln p_{min}^{n+1} \ =\ \frac{n+1}{n}\ln p_{min}\\ &\ge 2\ln p_{min}\,.\end{aligned}$$

Thus, $\lim_{n\to\infty}\frac{1}{n}\ln L_n \neq -\infty$. We have $\sup_{\alpha,\beta} N_{\alpha,\beta} \le N_\perp$. Also,

$$\bigcap_{\alpha,\beta}\left\{L_n \notin \left(\frac{\alpha}{1-\beta}, \frac{1-\alpha}{\beta}\right)\right\} = \{L_n = 0\}$$

for all $n \in \mathbb{N}$ and so $\sup_{\alpha,\beta} N_{\alpha,\beta} = N_\perp$. The final claim follows because

$$\begin{aligned}\lim_{\alpha,\beta\to 0} \mathbb{P}_{\pi_2}(N_{\alpha,\beta} < \infty) &= \mathbb{P}_{\pi_2}\left(\bigcap_{\alpha,\beta}\{N_{\alpha,\beta} < \infty\}\right) \\ &= \mathbb{P}_{\pi_2}\left(\sup_{\alpha,\beta} N_{\alpha,\beta} < \infty\right) \\ &= \mathbb{P}_{\pi_2}\left(N_\perp < \infty\right) \\ &= \mathbb{P}_{\pi_2}\left(\exists n \in \mathbb{N} : L_n = 0\right).\end{aligned}$$

We have shown that $\{L_n > 0\ \forall n \in \mathbb{N}\} \subseteq E^c_{-\infty}$. Thus, $E_{-\infty} \subseteq \{\exists n \in \mathbb{N} : L_n = 0\}$. The reverse inclusion is clear and so $\{N_\perp < \infty\} = E_{-\infty}$ holds which implies the last point in the proposition. □

Applying this to Example 6.3.4, we obtain $\sup_{\alpha,\beta} \mathbb{E}_{\pi_2}\left[N_{\alpha,\beta} \mid E_{-\infty}\right] = 3$.

6.3.3 Likelihood Exponent in $(-\infty, 0)$

Conditioned on E_ℓ where $\ell \in (-\infty, 0)$, Theorem 6.2.4 states that $N_{\alpha,\beta}$ scales with $\frac{\ln\alpha}{\ell}$ in expectation. The following result shows that this relationship also holds $\mathbb{P}_{\pi_2}$-almost surely.

Proposition 6.3.6. *Let $\ell \in \Lambda$ and assume $\ell \in (-\infty, 0)$. We have*

$$\mathbb{P}_{\pi_2}\left(N_{\alpha,\beta} \sim \frac{\ln\alpha}{\ell} \quad (as\ \alpha,\beta \to 0) \,\middle|\, E_\ell\right) \;=\; 1.$$

In fact, we prove the first part of theorem 6.2.4 using Proposition 6.3.6. If there were a bound $K \in \mathbb{N}$ such that $\mathbb{P}_{\pi_2}$-a.s. $\frac{N_{\alpha,\beta}}{-\ln\alpha} \le K$, the first part of Theorem 6.2.4 would follow from Proposition 6.3.6 by the dominated convergence theorem. However this is not the case in general. Instead we show in section 6.6 that the set of random variables $\{\frac{N_{\alpha,\beta}}{-\ln\alpha} \mid 0 < \alpha,\beta \le \frac{1}{2}\}$ is uniformly integrable with respect to the measure $\mathbb{P}_{\pi_2}$ and then use Vitali's convergence theorem.

Example 6.3.7. Recall Example 6.0.2, where $\Lambda = \{\ell\}$. Figure 6.2 demonstrates the asymptotic relationship in Proposition 6.3.6. Each of the 50 lines correspond to a sample run and we record the value of $N_{\alpha,\beta}$ for $0 \le -\ln\alpha = -\ln\beta \le 1000$. From the figure we estimate $-\frac{1}{\ell}$ as $\frac{10^5}{800} = 125$. This coincides with the estimate given in Example 6.2.1.

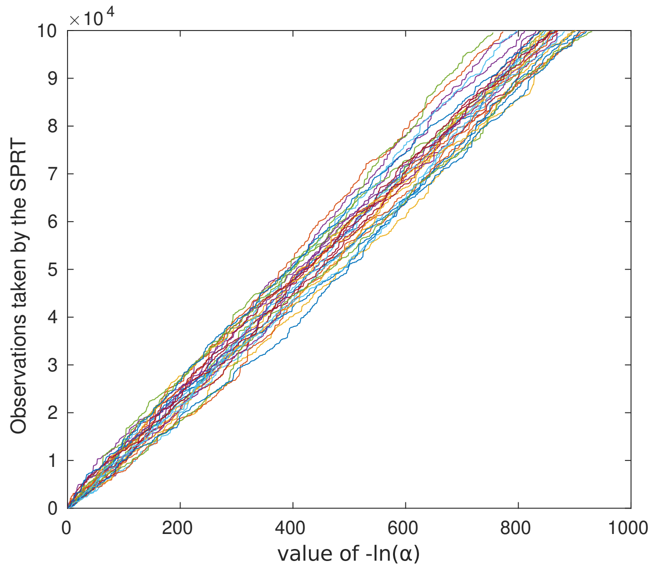


Figure 6.2: The number of observations taken by the SPRT for $0 \le -\ln\alpha = -\ln\beta \le 1000$.

We conclude from this section that the performance of the SPRT, in terms of its termination time $N_{\alpha,\beta}$, is tightly connected to likelihood exponents. This motivates our study of likelihood exponents in the next section.

6.4 Distribution of Likelihood Exponents

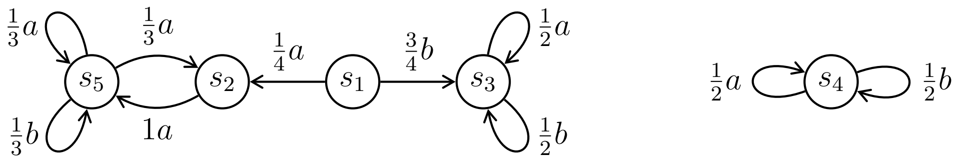
In this section we aim at computing $\mathbb{P}_{\pi_2}(E_\ell)$ for a likelihood exponent ℓ. We show the following theorem.

Theorem 6.4.1. *Given a HMM and initial distributions π_1, π_2,*

1. *one can compute $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n = -\infty)$ and $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n = 0)$ in polynomial space;*
2. *one can decide whether $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n = 0) = 0$ (i.e., $0 \notin \Lambda_{\pi_1,\pi_2}$) in polynomial time;*
3. *deciding whether $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n = 0) = 1$, whether $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n = -\infty) = 0$, and whether $\mathbb{P}_{\pi_2}(\lim_{n\to\infty} \frac{1}{n} \ln L_n = -\infty) = 1$ are all PSPACE-complete problems.*

The following example illustrates the construction underlying the PSPACE upper bound.

Example 6.4.2. Consider another adaption of Example 6.2.2.



If the first letter produced by s_4 is b, then $L_n = \frac{3}{2}$ for all $n \in \mathbb{N}$. If the first two letters are ab, then $L_1 = \frac{1}{2}$ and $L_n = 0$ for $n \geq 2$. If the first two letters are aa, then $s_5 \in \operatorname{supp}(e_{s_1} M(aaw))$ for all $w \in \Sigma^*$, and therefore, up to a $\mathbb{P}_{s_4}$-null set, $L_n > 0$ holds for all $n \in \mathbb{N}$ which implies (using proposition 6.3.5) that there is $\ell \in (-\infty, 0]$ such that $\lim_{n\to\infty} \frac{1}{n} \ln L_n = \ell$. By Lemma 6.0.1 we will have $\ell \in (-\infty, 0)$ if $d(e_{s_5}, e_{s_4}) = 1$ which we show in the following. Thus, $\Lambda_{s_1,s_4} = \{-\infty, \ell, 0\}$.

Let M' be the restriction of M to the ordered set of states $Q' = \{s_5, s_2, s_4\}$. That is, we discard the rows and columns corresponding to the states s_1 and s_3. Then,

$$M'(a) = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \text{ and } M'(b) = \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

It follows that

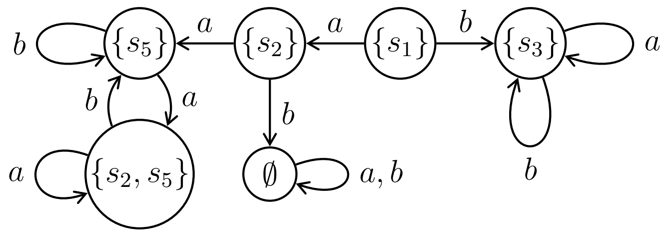
$$\begin{aligned} M'(a)\bar{1}^\top &= \left(\frac{2}{3}, 1, \frac{1}{2}\right)^\top \\ M'(b)\bar{1}^\top &= \left(\frac{1}{3}, 0, \frac{1}{2}\right)^\top \\ M'(aa)\bar{1}^\top &= \left(\frac{5}{9}, \frac{2}{3}, \frac{1}{4}\right)^\top. \end{aligned}$$

If there is $s, r \in \mathbb{R}$ such that $s\left(\frac{2}{3}, 1, \frac{1}{2}\right) + r\left(\frac{1}{3}, 0, \frac{1}{2}\right) = \left(\frac{5}{9}, \frac{2}{3}, \frac{1}{4}\right)$, then due to the second coordinate, $s = \frac{2}{3}$. Then the first coordinate implies $r = \frac{1}{3}$ but this creates a contradiction on the third coordinate since $\frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{2}$. It follows that the set $\left\{\left(\frac{2}{3}, 1, \frac{1}{2}\right)^\top, \left(\frac{1}{3}, 0, \frac{1}{2}\right)^\top, \left(\frac{5}{9}, \frac{2}{3}, \frac{1}{4}\right)^\top\right\}$ is linearly independent. Therefore,

$$\operatorname{span}\{M(w)\bar{1}^\top \mid w \in \Sigma^*\} = \mathbb{R}^3$$

and so by Proposition 3.1.1 for subdistributions μ_1, μ_2, $\mu_1 \equiv \mu_2$ implies $\mu_1 = \mu_2$. Further, $\operatorname{supp}(e_{s_5} M(w))$ is disjoint from $\operatorname{supp}(e_{s_4} M(w))$ for all $w \in \Sigma^*$ and so by Lemma 4.4.4 we have $d(e_{s_5}, e_{s_4}) = 1$.

The likelihood ratio L_n is 0 if and only if $\mathrm{supp}\,(\pi_1 M(w_n)) = \emptyset$. In order to track the support of $\pi_1 M(w_n)$, we consider the left part of the HMM as an NFA with s_1 as the initial state and its determinisation as shown in the DFA below.



Almost surely, s_4 produces a word that drives this DFA into a bottom SCC, which then determines $\lim_{n\to\infty} \frac{1}{n} \ln L_n$: concretely, the bottom SCC $\{\{s_5\}, \{s_2, s_5\}\}$ is associated with ℓ, the bottom SCC $\{\emptyset\}$ with $-\infty$, and the bottom SCC $\{\{s_3\}\}$ with 0.

In general, the observations need not be produced uniformly at random but by a HMM. Therefore, in the following construction refered to as a *Numerator Support Markov Chain*, we also keep track of the "current" state of the HMM which produces the observations. For $S \subseteq Q$ and $a \in \Sigma$, define $\delta(S, a) := \{q' \in Q \mid \exists\, q \in S : M(a)_{q,q'} > 0\}$. Define the Markov chain $\mathcal{D} := (2^Q \times Q, T)$ where

$$T_{(S,q),(S',q')} \;:=\; \sum_{\delta(S,a)=S'} M(a)_{q,q'}\,.$$

Given initial distributions π_1, π_2 on Q as before, define an initial distribution ι on $2^Q \times Q$ by $\iota((\mathrm{supp}\,(\pi_1), q)) := (\pi_2)_q$. Intuitively, the left part S of a state (S, q) tracks the support of $\pi_1 M(w_n)$, and the right part q tracks the current state of the HMM that had been initialised at a random state from π_2. The following lemma states the key properties of this construction.

Lemma 6.4.3. *Consider the Markov chain $\mathcal{D} = (2^Q \times Q, T)$ defined above.*

1. *Every bottom SCC of $\mathcal{D}$ is associated with a single likelihood exponent; i.e., for every bottom SCC $C \subseteq 2^Q \times Q$ there is $\ell(C) \in [-\infty, 0]$ such that for any initial distribution $\pi_1 \in [0,1]^Q$ and any state $q_2 \in Q$ with $(\mathrm{supp}\,(\pi_1), q_2) \in C$ we have $\Lambda_{\pi_1, e_{q_2}} = \{\ell(C)\}$.*

2. *Let $(S, q) \in C$ for a bottom SCC C. If $S = \emptyset$ then $\ell(C) = -\infty$; otherwise, if e_q and the uniform distribution on S are not distinguishable then $\ell(C) = 0$; otherwise $\ell(C) \in (-\infty, 0)$.*

3. *We have $\mathbb{P}_{\pi_2}(E_\ell) = \mathbb{P}_\iota(\{\text{visit bottom SCC } C \text{ with } \ell(C) = \ell\})$.*

All parts of the lemma rely on the observation that $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ depend only on the support of π_1 and on the support of π_2. The first part of the lemma follows from Lévy's 0-1 law. We use this lemma for the proof of the first part of theorem 6.4.1.

6.4.1 Proof of Theorem 6.4.1

Below we refer to the complexity class NC, the subclass of P comprising those problems solvable in polylogarithmic time by a parallel random-access machine using polynomially many processors; see, e.g., [38, Chapter 15]. To prove membership in PSPACE in a modular way, we use the following pattern:

Lemma 6.4.4. *Let P_1, P_2 be two problems, where P_2 is in NC. Suppose there is a reduction from P_1 to P_2 implemented by a PSPACE transducer, i.e., a Turing machine whose work tape (but not necessarily its output tape) is PSPACE-bounded. Then, P_1 is in PSPACE.*

Proof. Note that the output of the transducer is (at most) exponential. Problems in NC can be decided in polylogarithmic space [39, Theorem 4]. Using standard techniques for composing space-bounded transducers (see, e.g., [38, Proposition 8.2]), it follows that P_1 is in PSPACE. □

Towards a proof of the third point of theorem 6.4.1, we use the *mortality* problem, which asks, given a finite set of states Q, a finite alphabet Σ, and a function $\Phi : \Sigma \to \{0,1\}^{Q\times Q}$, whether there exists a word $w \in \Sigma^*$ such that $\Phi(w)$ is the zero matrix. The mortality problem can be viewed as a special case of the NFA non-universality problem (given an NFA, does it reject some word?). Like NFA universality, the mortality problem is PSPACE-complete [40].

Concerning $\mathbb{P}_{\pi_2}(E_{-\infty})$ (cf. the third point of theorem 6.4.1), we actually show a stronger result, namely that any nontrivial approximation of $\mathbb{P}_{\pi_2}(E_{-\infty})$ is PSPACE-hard. The proof is also based on the mortality problem.

Proposition 6.4.5. *There is a polynomial-time computable function that maps any instance of the mortality problem to a HMM and initial distributions π_1, π_2 so that if the instance is positive then $\mathbb{P}_{\pi_2}(E_{-\infty}) = 1$ and if the instance is negative then $\mathbb{P}_{\pi_2}(E_{-\infty}) = 0$. Thus, any nontrivial approximation of $\mathbb{P}_{\pi_2}(E_{-\infty})$ is PSPACE-hard.*

Proof. Let (Q, Σ, Φ) be an instance of the mortality problem. If there is $q \in Q$ that indexes a zero row in $\sum_{a\in\Sigma} \Phi(a)$, remove the row and column indexed by q in all $\Phi(a)$. Thus, we can assume without loss of generality that $\sum_{a\in\Sigma} \Phi(a)$ has no zero row. Construct a HMM (Q, Σ, M) so that $\Phi(a)$ and $M(a)$ have the same zero pattern for all $a \in \Sigma$. Define π_1 as a uniform distribution on Q. Define π_2 as a Dirac distribution on a fresh state that emits letters from Σ uniformly at random. Thus, if (Q, Σ, Φ) is a positive instance of the mortality problem then $\mathbb{P}_{\pi_2}(E_{-\infty}) = 1$, and if (Q, Σ, Φ) is a negative instance then $\mathbb{P}_{\pi_2}(E_{-\infty}) = 0$. □

Now we prove the main result of this section.

Proof of Theorem 6.4.1.

1. The Markov chain $\mathcal{D}$ from lemma 6.4.3 is exponentially big but can be constructed by a PSPACE transducer, i.e., a Turing machine whose work tape (but not necessarily its output tape) is PSPACE-bounded. The DAG (directed acyclic graph) structure, including the SCCs, of a graph can be computed in NL, which is included in NC. Using the pattern of lemma 6.4.4, the DAG structure of the Markov chain $\mathcal{D}$ can be computed in polynomial space. Thus, there is a PSPACE transducer that computes both $\mathcal{D}$ and its DAG structure.

   For each bottom SCC C, the PSPACE transducer also decides whether $\ell(C) = -\infty$ or $\ell(C) \in (-\infty, 0)$ or $\ell(C) = 0$, using the second point of lemma 6.4.3 and the polynomial-time algorithm for distinguishability from [23]. Finally, to compute $\mathbb{P}_{\pi_2}(E_{-\infty})$ and $\mathbb{P}_{\pi_2}(E_0)$, by the third point of lemma 6.4.3, it suffices to set up and solve a linear system of equations for computing hitting probabilities in a Markov chain. This system can also be computed by a PSPACE transducer. Linear systems of equations can be solved in NC [41, Theorem 5]. Using lemma 6.4.4 again, we conclude that one can compute $\mathbb{P}_{\pi_2}(E_{-\infty})$ and $\mathbb{P}_{\pi_2}(E_0)$ in polynomial space.

2. Immediate from corollary 6.3.3 and the polynomial-time decidability of distinguishability (Theorem 4.0.1).

3. The claims concerning $\mathbb{P}_{\pi_2}(E_{-\infty})$ follow from part 1 and proposition 6.4.5. Consider the problem whether $\mathbb{P}_{\pi_2}(E_0) = 1$. By part 1, it is in PSPACE. Towards PSPACE-hardness we reduce again from mortality. Let (Q, Σ, Φ) be an instance of the mortality problem. Let $Q' := Q \cup \{q_\perp, q_2\}$ for fresh states $q_\perp, q_2$, and let $\Sigma' := \Sigma \cup \{\$\}$ for a fresh letter \$. Obtain Φ' from Φ by adding, for every $q \in Q'$, a \$-labelled transition to $q_\perp$, and an a-labelled loop from q_2 to itself for all $a \in \Sigma$. Construct a HMM (Q', Σ', M) so that $\Phi'(a)$ and $M(a)$ have the same zero pattern for all $a \in \Sigma'$ (e.g., use uniform distributions). See fig. 6.3.

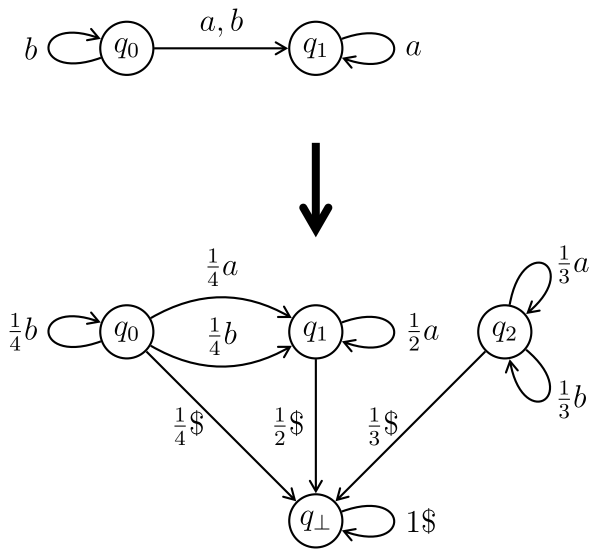


Figure 6.3: Illustration of the reduction from mortality to $\mathbb{P}_{\pi_2}(E_0) < 1$. In this example, $\Phi(ab)$ is the zero matrix. Accordingly, we have $\mathbb{P}_{\pi_2}(E_0) < 1$, as $L_2(abw) = 0$ for all $w \in \Sigma^\omega$.

Let $\pi_1 \in [0,1]^{Q'}$ be the uniform distribution on Q (i.e., $(\pi_1)_{q_\perp} = (\pi_1)_{q_2} = 0$), and let π_2 be the Dirac distribution on q_2.

Suppose (Q, Σ, Φ) is a positive instance of the mortality problem. Let $v \in \Sigma^*$ such that $\Phi(v)$ is the zero matrix. Then $L_{|v|}(vw) = 0$ holds for all $w \in \Sigma^\omega$. It follows that $\mathbb{P}_{\pi_2}(E_{-\infty}) > 0$ and so $\mathbb{P}_{\pi_2}(E_0) < 1$.

Conversely, suppose (Q, Σ, Φ) is a negative instance of the mortality problem. The word produced from q_2 contains $\mathbb{P}_{e_{q_2}}$-a.s. the letter \$, i.e., is of the form $u\$v$ for $u \in \Sigma^*$ and $v \in (\Sigma \cup \{\$\})^\omega$. Since (Q, Σ, Φ) is a negative instance, it follows that $\mathrm{supp}\,(\pi_1 M(u\$)) = \{q_\perp\} = \mathrm{supp}\,(e_{q_2} M(u\$))$. Thus, $\lim_{n\to\infty} L_n > 0$. Hence, $\mathbb{P}_{\pi_2}(E_0) = 1$. □

6.5 The Numerator Support Markov Chain

This section is dedicated to the proof of lemma 6.4.3 which we restate and prove now.

Lemma 6.4.3. *Consider the Markov chain $\mathcal{D} = (2^Q \times Q, T)$ defined above.*

1. *Every bottom SCC of $\mathcal{D}$ is associated with a single likelihood exponent; i.e., for every bottom SCC $C \subseteq 2^Q \times Q$ there is $\ell(C) \in [-\infty, 0]$ such that for any initial distribution $\pi_1 \in [0,1]^Q$ and any state $q_2 \in Q$ with $(\mathrm{supp}\,(\pi_1), q_2) \in C$ we have $\Lambda_{\pi_1, e_{q_2}} = \{\ell(C)\}$.*
2. *Let $(S, q) \in C$ for a bottom SCC C. If $S = \emptyset$ then $\ell(C) = -\infty$; otherwise, if e_q and the uniform distribution on S are not distinguishable then $\ell(C) = 0$; otherwise $\ell(C) \in (-\infty, 0)$.*
3. *We have $\mathbb{P}_{\pi_2}(E_\ell) = \mathbb{P}_\iota(\{$visit bottom SCC C with $\ell(C) = \ell\})$.*

Proof. 1. Let $C \subseteq 2^Q \times Q$ be a bottom SCC of $\mathcal{D}$. Let π, π' be distributions on Q and $q, q' \in Q$ such that $(\mathrm{supp}\,(\pi), q), (\mathrm{supp}\,(\pi'), q') \in C$. Suppose that $\ell \in \Lambda_{\pi, e_q}$; i.e.,

$$\mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi M(w_n)\|}{\|e_q M(w_n)\|} = \ell\right) \;=\; x \quad \text{for some } x > 0. \tag{6.3}$$

It suffices to show that $\Lambda_{\pi', e_{q'}} = \{\ell\}$, i.e.,

$$\mathbb{P}_{e_{q'}}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi' M(w_n)\|}{\|e_{q'} M(w_n)\|} = \ell\right) \;=\; 1.$$

By Lévy's 0-1 law it suffices to show that for all paths $q' a_1 q_1 \cdots a_m q_m$ with $\mathbb{P}_{e_{q'}}(q' a_1 q_1 \cdots a_m q_m (\Sigma Q)^\omega) > 0$ there is $y > 0$ with

$$\mathbb{P}_{e_{q'}}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi' M(w_n)\|}{\|e_{q'} M(w_n)\|} = \ell \;\middle|\; q' a_1 q_1 \cdots a_m q_m (\Sigma Q)^\omega\right) \;\geq\; y. \tag{6.4}$$

Let $u = q'a_1q_1 \cdots a_mq_m$ be a path with $\mathbb{P}_{e_{q'}}(u(\Sigma Q)^\omega) > 0$. Since C is a bottom SCC of $\mathcal{D}$, we have

$$\mathbb{P}_{e_{q'}}\left(\exists\, k \geq m : \mathrm{supp}\,(\pi' M(a_1 \cdots a_m \cdots a_k)) = \mathrm{supp}\,(\pi),\ q_k = q \mid u(\Sigma Q)^\omega\right) = 1\,.$$

Thus, letting $v = q'a_1q_1 \cdots a_mq_m \cdots a_kq_k$, with $k \geq m$, be an arbitrary extension of u with $\mathbb{P}_{e_{q'}}(v(\Sigma Q)^\omega) > 0$ and $\mathrm{supp}\,(\pi' M(a_1 \cdots a_m \cdots a_k)) = \mathrm{supp}\,(\pi)$ and $q_k = q$, we have $\mathbb{P}_{e_{q'}}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi' M(w_n)\|}{\|e_{q'} M(w_n)\|} = \ell \;\middle|\; v(\Sigma Q)^\omega\right) > 0$ implies $\mathbb{P}_{e_{q'}}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi' M(w_n)\|}{\|e_{q'} M(w_n)\|} = \ell \;\middle|\; u(\Sigma Q)^\omega\right) > 0$. Further,

$$\begin{aligned}
&\mathbb{P}_{e_{q'}}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi' M(w_n)\|}{\|e_{q'} M(w_n)\|} = \ell \;\middle|\; v(\Sigma Q)^\omega\right) \\
&\geq \mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(\pi' M(a_1 \cdots a_k))M(w_n)\|}{\|(e_{q'} M(a_1 \cdots a_k))M(w_n)\|} = \ell\right) \\
&\geq \mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(\pi' M(a_1 \cdots a_k))M(w_n)\|}{\|e_q M(w_n)\|} = \ell \ \text{ and} \right. && (6.5)\\
&\qquad \left. \lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(e_{q'} M(a_1 \cdots a_k))M(w_n)\|}{\|e_q M(w_n)\|} = 0\right) && (6.6)
\end{aligned}$$

Concerning the event in (6.5), by (6.3) and since $\mathrm{supp}\,(\pi' M(a_1 \cdots a_k)) = \mathrm{supp}\,(\pi)$, we have

$$\mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(\pi' M(a_1 \cdots a_k))M(w_n)\|}{\|e_q M(w_n)\|} = \ell\right) \ \geq\ x\,.$$

Concerning the event in (6.6), it follows from the first point of lemma 6.0.1 that

$$\mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(e_{q'} M(a_1 \cdots a_k))M(w_n)\|}{\|e_q M(w_n)\|} \leq 0\right) \ =\ 1\,.$$

Further, since $\mathbb{P}_{e_{q'}}(q'a_1q_1 \cdots a_kq_k(\Sigma Q)^\omega) > 0$, we have $q \in \mathrm{supp}\,(e_{q'} M(a_1 \cdots a_k))$ and so

$$\mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(e_{q'} M(a_1 \cdots a_k))M(w_n)\|}{\|e_q M(w_n)\|} \geq 0\right) \ =\ 1\,.$$

Thus, continuing the inequality chain from above, we conclude that

$$\mathbb{P}_{e_{q'}}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi' M(w_n)\|}{\|e_{q'} M(w_n)\|} = \ell \;\middle|\; u(\Sigma Q)^\omega\right) \ \geq\ x\,,$$

proving (6.4), as desired.

2. Let $(S,q) \in C$ for a bottom SCC C. If $S = \emptyset$ then we may define $\ell(C) = -\infty$. Otherwise, let π_S denote the uniform distribution on S. Suppose that π_S and e_q are not distinguishable. By corollary 6.3.3 it follows that $0 \in \Lambda_{\pi_S, e_q}$. Using part 1 we obtain $\ell(C) = 0$. Finally, suppose that π_S and e_q are distinguishable. By corollary 6.3.3 it follows that $0 \notin \Lambda_{\pi_S, e_q}$. Since C does not contain any states of the form $(\emptyset, q')$, by proposition 6.3.5 we have $-\infty \notin \Lambda_{\pi_S, e_q}$. Using part 1 we obtain $\ell(C) \in (-\infty, 0)$.

3. We define a function f that maps paths of $\mathcal{F}$ to paths of $\mathcal{D}$ as follows. Set $f(q_0 a_1 q_2 \cdots a_m q_m) := (S_0, q_0)(S_1, q_1) \cdots (S_m, q_m)$ where $S_0 = \mathrm{supp}\,(\pi_1)$ and $\delta(S_{i-1}, a_i) = S_i$ for all $1 \le i \le m$. The Markov chain $\mathcal{D}$ is constructed so that for any path $v = (S_0, q_0)(S_1, q_1) \cdots (S_m, q_m)$ we have

$$\mathbb{P}_\iota(v(2^Q \times Q)^\omega) \;=\; \mathbb{P}_{\pi_2}(f^{-1}(v)(\Sigma Q)^\omega)\,.$$

Let C be any bottom SCC, and let $\ell = \ell(C)$. Define the event

$$V_C := \{q_0 a_1 q_1 \cdots \in Q(\Sigma Q)^\omega \mid \exists\, m \in \mathbb{N} : f(q_0 a_1 q_1 \cdots a_m q_m) \text{ ends in } C\}\,.$$

So we have $\mathbb{P}_\iota(\{\text{visit } C\}) = \mathbb{P}_{\pi_2}(V_C)$, and it suffices to show that $\mathbb{P}_{\pi_2}(E_\ell \mid V_C) = 1$. Let $u = q_0 a_1 q_1 \cdots a_m q_m$ be a path with $\mathbb{P}_{\pi_2}(u(\Sigma Q)^\omega) > 0$ such that $f(u)$ ends in C, say in $(S,q) \in C$, with $q = q_m$. Thus, $\mathrm{supp}\,(\pi_1 M(a_1 \cdots a_m)) = S$ and $q \in \mathrm{supp}\,(\pi_2 M(a_1 \cdots a_m))$. It suffices to show that $\mathbb{P}_{\pi_2}(E_\ell \mid u(\Sigma Q)^\omega) = 1$. We have:

$$\begin{aligned}
&\mathbb{P}_{\pi_2}(E_\ell \mid u(\Sigma Q)^\omega)\\
&= \mathbb{P}_{\pi_2}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi_1 M(w_n)\|}{\|\pi_2 M(w_n)\|} = \ell \;\middle|\; u(\Sigma Q)^\omega\right)\\
&= \mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(\pi_1 M(a_1 \cdots a_m))M(w_n)\|}{\|(\pi_2 M(a_1 \cdots a_m))M(w_n)\|} = \ell\right)\\
&\ge \mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(\pi_1 M(a_1 \cdots a_m))M(w_n)\|}{\|e_q M(w_n)\|} = \ell \;\text{ and}\right. && (6.7)\\
&\qquad\left.\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(\pi_2 M(a_1 \cdots a_m))M(w_n)\|}{\|e_q M(w_n)\|} = 0\right) && (6.8)
\end{aligned}$$

Concerning the event in (6.7), by part 2 and since $\mathrm{supp}\,(\pi_1 M(a_1 \cdots a_m)) = S$, we have

$$\mathbb{P}_{e_q}\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(\pi_1 M(a_1 \cdots a_m))M(w_n)\|}{\|e_q M(w_n)\|} = \ell\right) \;=\; 1\,.$$

Concerning the event in (6.8), it follows from the first point of lemma 6.0.1 that

$$\mathbb{P}_{e_q}\left(\lim_{n\to\infty}\frac{1}{n}\ln\frac{\|(\pi_2 M(a_1\cdots a_m))M(w_n)\|}{\|e_q M(w_n)\|}\leq 0\right) \;=\; 1\,.$$

Further, since $q \in \mathrm{supp}\,(\pi_2 M(a_1\cdots a_m))$, we have

$$\mathbb{P}_{e_q}\left(\lim_{n\to\infty}\frac{1}{n}\ln\frac{\|(\pi_2 M(a_1\cdots a_m))M(w_n)\|}{\|e_q M(w_n)\|}\geq 0\right) \;=\; 1\,.$$

Thus, the events in (6.7) and (6.8) occur $\mathbb{P}_{e_q}$-a.s. We conclude that $\mathbb{P}_{\pi_2}(E_\ell \mid u(\Sigma Q)^\omega) = 1$, as desired. □

6.6 A Generalised Wald Formula

This section is dedicated to the proof of Theorem 6.2.4. We begin by restating the following.

Proposition 6.3.2. *The following two equalities hold up to $\mathbb{P}_{\pi_2}$-null sets:*

$$E_0 \;=\; \left\{\lim_{n\to\infty} L_n > 0\right\} \;=\; \bigcup_{\alpha,\beta>0}\{N_{\alpha,\beta}=\infty\}\,.$$

Thus, $\lim_{\alpha,\beta\to 0}\mathbb{P}_{\pi_2}(N_{\alpha,\beta}=\infty) = \mathbb{P}_{\pi_2}(E_0)$.

Towards the proof of Proposition 6.3.2 we use the following which is Theorem 5 from [24].

Lemma 6.6.1. *Let (Q,Σ,M) be a HMM and let π_1 and π_2 be initial distributions. If π_1 and π_2 are distinguishable then there is $c>0$ such that*

$$\mathbb{P}_{\pi_2}\Big(L_{2|Q|n}\leq 1\Big) - \mathbb{P}_{\pi_1}\Big(L_{2|Q|n}\geq 1\Big) \geq 1 - 2\exp\Big(-\frac{c^2}{18}n\Big).$$

Proof of Proposition 6.3.2. Towards the first equality, if $\lim_{n\to\infty} L_n > 0$ then $\lim_{n\to\infty}\frac{1}{n}\ln L_n = 0$ and so we only need to show that $E_0 \subseteq \{\lim_{n\to\infty} L_n > 0\}$. Assume, $0\in\Lambda$. By lemma 6.4.3 there are a set of bottom SCCs $\mathcal{Z}$ in $\mathcal{D}$ such that for some $Z\in\mathcal{Z}$ we have $\ell(Z)=\{0\}$. Let $\pi\in[0,1]^Q$ and $r\in Q$ such that $(\mathrm{supp}\,(\pi), r)\in Z$. Suppose that π and δ_r are distinguishable then by Lemma 6.6.1 we have $\mathbb{P}_{\delta_r}(\tilde{L}_n \geq 1) \leq 2\exp\Big(-\frac{c^2}{18}n\Big)$ and $\mathbb{P}_{\pi}(\tilde{L}_n \leq 1) \leq 2\exp\Big(-\frac{c^2}{18}n\Big)$ for all

$n \in \mathbb{N}$ where $\tilde{L}_n$ is the likelihood ratio started from initial distributions π and δ_r. Fix $-\frac{c^2}{18} < \alpha \leq 0$ and define the event $W_n = \{1 > \tilde{L}_n \geq \mathrm{e}^{n\alpha}\}$. Then,

$$\begin{aligned}
\mathbb{P}_{\delta_r}\left(\lim_{n\to\infty}\frac{1}{n}\ln\tilde{L}_n > \alpha\right) &\leq \mathbb{P}_{\delta_r}\left(\liminf_n\left\{\frac{1}{n}\ln\tilde{L}_n \geq \alpha\right\}\right) \\
&\leq \liminf_n \mathbb{P}_{\delta_r}\left(\frac{1}{n}\ln\tilde{L}_n \geq \alpha\right) \\
&\leq \liminf_n \mathbb{P}_{\delta_r}(\tilde{L}_n \geq \mathrm{e}^{n\alpha}) \\
&= \liminf_n \left[\mathbb{P}_{\delta_r}(1 > \tilde{L}_n \geq \mathrm{e}^{n\alpha}) + \mathbb{P}_{\pi_2}(\tilde{L}_n \geq 1)\right] \\
&\leq \liminf_n \left[\sum_{w\in W_n}\delta_r M(w)\overline{1}^\top + 2\exp\Big(-\frac{c^2}{18}n\Big)\right] \\
&\leq \liminf_n \left[e^{-n\alpha}\sum_{w\in W_n}\pi M(w)\overline{1}^\top\right] \\
&\leq \liminf_n \left[e^{-n\alpha}\mathbb{P}_\pi(\tilde{L}_n < 1)\right] \\
&\leq \liminf_n \left[2e^{-n\alpha}\exp\Big(-\frac{c^2}{18}n\Big)\right] \\
&= 0.
\end{aligned}$$

In particular, $\mathbb{P}_{\pi_2}(\lim_{n\to\infty}\frac{1}{n}\ln\tilde{L}_n = 0) = 0$ which contradicts $0 \in \Lambda$. Hence π and δ_r are not distinguishable and so $\mathbb{P}_{\delta_r}$-almost surely, we have $\lim_{n\to\infty}\tilde{L}_n > 0$. By conditioning on the events

$$\{w = a_1 r_1 \cdots a_n r_n \in (\Sigma Q)^* \mid \operatorname{supp}(\pi_1 M(w)) = \operatorname{supp}(\pi), r_n = r\}$$

it follows that $\mathbb{P}_{\pi_2}(E_0) = \mathbb{P}_{\pi_2}(\{\lim_{n\to\infty} L_n > 0\})$.

We now show the second equality. If $\lim_{n\to\infty} L_n > 0$ then for α, β small enough, L_n never crosses the SPRT bounds. Hence, we have $\{\lim_{n\to\infty} L_n > 0\} \subseteq \bigcup_{\alpha,\beta}\{N_{\alpha,\beta} = \infty\}$. For the converse inclusion, suppose that $N_{\alpha,\beta} = \infty$ for some α, β. This would contradict $\lim_{n\to\infty} L_n = 0$ since then $N_{\alpha,\beta}$ would be $\mathbb{P}_{\pi_2}$-almost surely finite. □

We now state and prove Proposition 6.3.6 which serves as an intermediate step towards Theorem 6.2.4.

Proposition 6.3.6. *Let $\ell \in \Lambda$ and assume $\ell \in (-\infty, 0)$. We have*

$$\mathbb{P}_{\pi_2}\Big(N_{\alpha,\beta} \sim \frac{\ln\alpha}{\ell} \quad (\text{as } \alpha, \beta \to 0) \;\Big|\; E_\ell\Big) \;=\; 1.$$

Proof. Fix $n \in \mathbb{N}$ and $w \in E_\ell$. Since $\ell > -\infty$ we may assume $L_n(w) > 0$ and so $\pi_1 M(w_n)$ contains at least one positive entry. Let $p_{min} > 0$ be the minimum non-zero entry in π_1, π_2 and all $M(a)$ where $a \in \Sigma$. Then, $\pi_1 M(w_n) \geq p_{min}^{n+1}$. We have $\pi_2 M(w_n) \leq 1$ and so $L_n(w) \geq p_{min}^{n+1}$. A similar argument shows $L_n(w) \leq p_{min}^{-(n+1)}$.

We now relax the dependence on n. Since $\ell \in (-\infty, 0)$ we have that $\ln L_n(w) \to -\infty$. Further, we have the following

$$\begin{aligned}
& L_{N_{\alpha,\beta}}(w) \leq \frac{\alpha}{1-\beta} \quad \text{or} \quad L_{N_{\alpha,\beta}}(w) \geq \frac{\beta}{1-\alpha} \\
\implies & \pi_1^{\min} M_{\min}^{N_{\alpha,\beta}(w)} \leq \frac{\alpha}{1-\beta} \quad \text{or} \quad \pi_2^{\min} M_{\min}^{-N_{\alpha,\beta}(w)} \geq \frac{\beta}{1-\alpha} \\
\implies & N_{\alpha,\beta}(w) \geq \frac{1}{\ln M_{\min}} \ln\left(\frac{1}{\pi_1^{\min}} \frac{\alpha}{1-\beta}\right) \text{ or } N_{\alpha,\beta}(w) \geq \frac{1}{\ln M_{\min}} \ln\left(\frac{1}{\pi_2^{\min}} \frac{\beta}{1-\alpha}\right) \\
\implies & N_{\alpha,\beta}(w) \geq \frac{1}{\ln M_{\min}} \ln\left[\min\left(\frac{1}{\pi_1^{\min}} \frac{\alpha}{1-\beta}, \frac{1}{\pi_2^{\min}} \frac{\beta}{1-\alpha}\right)\right].
\end{aligned}$$

Hence, $N_{\alpha,\beta}(w) \to \infty$ as $\alpha, \beta \to 0$.

Let $U_{\alpha,\beta} = \{u \in E_\ell \mid \ln L_{N_{\alpha,\beta}(u)}(u) \leq \ln \frac{\alpha}{1-\beta}\}$. It follows that if $w \in E_\ell \setminus U_{\alpha,\beta}$ then $L_n(w)$ crosses the upper threshold before the lower threshold. Therefore, $w \in \bigcap_{\alpha,\beta \in (0,1]} E_\ell \setminus U_{\alpha,\beta}$ implies that $L_n(w)$ is unbounded in n. We have $L_n(w) \to 0$ since $\ell \in (-\infty, 0)$ and thus $\lim_{\alpha,\beta\to 0} \chi_{U_{\alpha,\beta}}(w) = 1$. Further,

$$\begin{aligned}
0 &\leq \chi_{U_{\alpha,\beta}} \frac{\ln \frac{\alpha}{1-\beta} - \ln L_{N_{\alpha,\beta}(w)}(w)}{N_{\alpha,\beta}(w)} \quad \text{since} \\
&\leq \chi_{U_{\alpha,\beta}} \frac{\ln L_{N_{\alpha,\beta}(w)-1}(w) - \ln L_{N_{\alpha,\beta}(w)}(w)}{N_{\alpha,\beta}(w)} \quad \text{since } \ln L_{N_{\alpha,\beta}(w)-1}(w) \geq \ln \frac{\alpha}{1-\beta} \\
&\to 0 \quad \text{as } \alpha, \beta \to 0.
\end{aligned}$$

Finally,

$$\begin{aligned}
\lim_{\alpha,\beta\to 0} \frac{\ln \alpha}{N_{\alpha,\beta}(w)} &= \lim_{\alpha,\beta\to 0} \chi_{U_{\alpha,\beta}} \frac{\ln \alpha}{N_{\alpha,\beta}(w)} \\
&= \lim_{\alpha,\beta\to 0} \chi_{U_{\alpha,\beta}} \frac{\ln \frac{\alpha}{1-\beta}}{N_{\alpha,\beta}(w)} \\
&= \lim_{\alpha,\beta\to 0} \chi_{U_{\alpha,\beta}} \frac{\ln L_{N_{\alpha,\beta}(w)}(w)}{N_{\alpha,\beta}(w)} \\
&= \lim_{\alpha,\beta\to 0} \frac{\ln L_{N_{\alpha,\beta}(w)}(w)}{N_{\alpha,\beta}(w)} \\
&= \ell.
\end{aligned}$$

□

Towards a proof of the first point of Theorem 6.2.4, we would like to show that the conditional almost sure convergence of $\frac{\ln \alpha}{N_{\alpha,\beta}(w)}$ as $\alpha, \beta \to \infty$ also happens in expectation. We restate the Theorem here:

Theorem 6.2.4 (Generalised Wald Formula)**.** *Let ℓ be a likelihood exponent and let π_1 and π_2 be initial distributions.*

1. *If $\ell \in (-\infty, 0)$ then $\mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] \sim \dfrac{\ln \alpha}{\ell}$ (as $\alpha, \beta \to 0$).*

2. *If $\ell = 0$ then there exist $\alpha, \beta > 0$ such that $\mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] = \infty$.*

3. *If $\ell = -\infty$ then $\sup\limits_{\alpha,\beta} \mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] < \infty$.*

Let Φ be a set of random variables on measure space $(\Sigma^\omega, \mathscr{G}^\omega, \mathbb{P}_{\pi_2})$ such that $\mathbb{E}_{\pi_2}[\phi] < \infty$ for all $\phi \in \Phi$. We say that Φ is *uniformly integrable* if

$$\lim_{K \to \infty} \sup_{\phi \in \Phi} \mathbb{E}_{\pi_2}\left[\phi \chi_{\phi \geq K}\right] = 0.$$

We use the Lebesgue-Vitali theorem, a proof of which can found on page 268 of [42].

Theorem 6.6.2 (Lebesgue-Vitali Theorem)**.** *Suppose ϕ is a random variable and $(\phi_n)_{n=1}^\infty$ is a sequence of random variables on measure space $(\Sigma^\omega, \mathscr{G}^\omega, \mathbb{P}_{\pi_2})$ such that $\mathbb{E}_{\pi_2}[\phi_n] < \infty$ for all $n \in \mathbb{N}$. The following assertions are equivalent:*

1. *The sequence $(\phi_n)_{n=1}^\infty$ converges to ϕ in $\mathbb{P}_{\pi_2}$-measure and is uniformly integrable;*

2. *We have $\mathbb{E}_{\pi_2}[\phi] < \infty$ and $\lim_{n\to\infty} \mathbb{E}_{\pi_2}[|\phi_n - \phi|] = 0$.*

By the Lebesgue-Vitali Theorem, we need to show that $\{\frac{N_{\alpha,\beta}}{-\ln \alpha} \mid 0 < \alpha, \beta \leq \frac{1}{2}\}$ is uniformly integrable. For this, we use another technical result which is Lemma 9 from [24].

Lemma 6.6.3. *There is a number $c > 0$, computable in polynomial time, such that*

$$\mathbb{P}_{\pi_2}\Big(L_{2|Q|n} \geq \exp\left(-\frac{c^2}{36}n\right)\Big) \leq 4\exp\Big(-\frac{c^2}{36}n\Big).$$

Proof of Theorem 6.2.4. Let $\mathbb{P}_\ell$ be the $\mathbb{P}_{\pi_2}$ measure conditioned on E_ℓ. By Proposition 6.3.6, we have $\lim_{\alpha,\beta\to 0}\frac{N_{\alpha,\beta}}{\ln\alpha}$ exists $\mathbb{P}_\ell$-almost surely. Hence, the convergence is also in $\mathbb{P}_\ell$-measure. Therefore, by the Lebesgue-Vitali Theorem (Theorem 6.6.2), it is sufficient to show that the set of random variables $\{\frac{N_{\alpha,\beta}}{\ln\alpha}\mid \alpha,\beta\in(0,\frac{1}{2})\}$ is uniformly integrable conditioned on E_ℓ. In fact, because

$$\lim_{K\to\infty}\sup_{\alpha,\beta}\mathbb{E}_{\pi_2}\left[\frac{N_{\alpha,\beta}}{-\ln\alpha}\chi_{\frac{N_{\alpha,\beta}}{-\ln\alpha}\geq K}\right]\geq \mathbb{P}_{\pi_2}(E_\ell)\lim_{K\to\infty}\sup_{\alpha,\beta}\frac{1}{-\ln\alpha}\mathbb{E}_{\pi_2}\left[\frac{N_{\alpha,\beta}}{-\ln\alpha}\chi_{\frac{N_{\alpha,\beta}}{-\ln\alpha}\geq K}\mid E_\ell\right],$$

it is sufficient to check the uniform integrability condition without conditioning on E_ℓ.

Fix $K\in\mathbb{N}$ and $\alpha,\beta\in(0,\frac{1}{2}]$. Since the random variable $N_{\alpha,\beta}\chi_{\frac{N_{\alpha,\beta}}{-\ln\alpha}\geq K}$ takes values in $\mathbb{N}_0$ we may write its expectation as a sum of probabilities:

$$\begin{aligned}
&\mathbb{E}_{\pi_2}\left[\frac{N_{\alpha,\beta}}{-\ln\alpha}\chi_{\frac{N_{\alpha,\beta}}{-\ln\alpha}\geq K}\right]\\
=&\frac{1}{-\ln\alpha}\mathbb{E}_{\pi_2}\left[N_{\alpha,\beta}\chi_{\frac{N_{\alpha,\beta}}{-\ln\alpha}\geq K}\right]\\
=&\frac{1}{-\ln\alpha}\sum_{n=0}^{\infty}\mathbb{P}_{\pi_2}(N_{\alpha,\beta}\chi_{\frac{N_{\alpha,\beta}}{-\ln\alpha}\geq K}>n)\\
=&\frac{1}{-\ln\alpha}\left[\lfloor -K\ln\alpha\rfloor\mathbb{P}_{\pi_2}(N_{\alpha,\beta}>\lfloor -K\ln\alpha\rfloor)+\sum_{n=\lfloor -K\ln\alpha\rfloor}^{\infty}\mathbb{P}_{\pi_2}(N_{\alpha,\beta}>n)\right].
\end{aligned}\tag{6.9}$$

We tackle each summand of the expression (6.9) individually. Let $m_\alpha=\left\lfloor\frac{-K\ln\alpha}{2|Q|}\right\rfloor$ then we have

$$\begin{aligned}
\frac{\lfloor -K\ln\alpha\rfloor}{-\ln\alpha}\mathbb{P}_{\pi_2}(N_{\alpha,\beta}>\lfloor -K\ln\alpha\rfloor)&\leq K\mathbb{P}_{\pi_2}(N_{\alpha,\beta}>2|Q|m_\alpha)\\
&\leq K\mathbb{P}_{\pi_2}\left(L_{2|Q|m_\alpha}\geq\frac{\alpha}{1-\beta}\right)
\end{aligned}$$

because $N_{\alpha,\beta}>2|Q|m_\alpha$ implies that the likelihood ratio has not crossed the lower threshold after $2|Q|m_\alpha$ observations. We have $m_\alpha\geq\frac{-K\ln\alpha}{4|Q|}$ for K large enough and α small enough. Taking c from Lemma 6.6.3 we have for large K:

$$\exp(-\frac{c^2}{36}m_\alpha)\leq\exp\left(\frac{c^2K\ln\alpha}{144|Q|}\right)\leq\alpha^{\frac{c^2K}{144|Q|}}\leq\alpha\leq\frac{\alpha}{1-\beta}.$$

This implies

$$K\mathbb{P}_{\pi_2}\left(L_{2|Q|m_\alpha} \geq \frac{\alpha}{1-\beta}\right) \leq K\mathbb{P}_{\pi_2}\left(L_{2|Q|m_\alpha} \geq \exp(-\frac{c^2}{36}m_\alpha)\right)$$
$$\leq 4K\exp\left(-\frac{c^2}{36}m_\alpha\right)$$
$$\leq 4K\exp\left(-\frac{c^2}{36}\left\lfloor\frac{K\ln 2}{2|Q|}\right\rfloor\right). \tag{6.10}$$

Thus $\lim_{K\to\infty}\sup_{\alpha,\beta} 4K\exp\left(-\frac{c^2}{36}m_\alpha\right) = \lim_{K\to\infty} 4K\exp\left(-\frac{c^2}{36}\left\lfloor\frac{K\ln 2}{2|Q|}\right\rfloor\right) = 0$.

The second part of the summand (6.9) proceeds similarly. Since $\mathbb{P}_{\pi_2}(N_{\alpha,\beta} > n)$ is monotonically decreasing as n increases, for all integer $m \geq m_\alpha$ and $2|Q|m \leq n \leq 2|Q|(m+1)$ we have $\mathbb{P}_{\pi_2}(N_{\alpha,\beta} > n) \leq \mathbb{P}_{\pi_2}(N_{\alpha,\beta} > 2|Q|m)$. Hence, with the same c from Lemma 6.6.3:

$$\frac{1}{-\ln\alpha}\sum_{n=\lfloor -K\ln\alpha\rfloor}^{\infty}\mathbb{P}_{\pi_2}(N_{\alpha,\beta} > n) \leq \frac{2|Q|}{-\ln\alpha}\sum_{m=m_\alpha}^{\infty}\mathbb{P}_{\pi_2}(N_{\alpha,\beta} > 2|Q|m)$$
$$\leq \frac{2|Q|}{-\ln\alpha}\sum_{m=m_\alpha}^{\infty}\mathbb{P}_{\pi_2}(L_{2|Q|m} \geq \alpha)$$
$$\leq \frac{2|Q|}{-\ln\alpha}\sum_{m=m_\alpha}^{\infty}\mathbb{P}_{\pi_2}\left(L_{2|Q|m} \geq \exp\left(-\frac{c^2}{36}m_\alpha\right)\right)$$
$$\leq \frac{8|Q|}{-\ln\alpha}\sum_{m=m_\alpha}^{\infty}\exp\left(-\frac{c^2}{36}m_\alpha\right) \text{ by Lemma 6.6.3}$$
$$= \frac{8|Q|}{-\ln\alpha}\times\frac{\exp-\frac{c^2}{36}m_\alpha}{1-\exp c^2/36}$$
$$\leq \frac{8|Q|}{-\ln 2}\times\frac{\exp-\frac{c^2}{36}\left(1+\frac{K\ln 2}{2|Q|}\right)}{1-\exp c^2/36}$$
$$\to 0 \text{ as } K\to\infty. \tag{6.11}$$

By combining (6.10) and (6.11) it follows from (6.9) that

$$\lim_{K\to\infty}\sup_{\alpha,\beta\in(0,\frac{1}{2}]}\mathbb{E}_{\pi_2}\left[\frac{N_{\alpha,\beta}}{-\ln\alpha}\chi_{\frac{N_{\alpha,\beta}}{-\ln\alpha}\geq K}\right] = 0. \tag{6.12}$$

Since (6.12) holds we may apply the reverse triangle inequality to yield the first point in the theorem. Towards the second point, we have $\mathbb{P}_{\pi_2}(E_0) > 0$. From Proposition 6.3.2 we have that $\mathbb{P}_{\pi_2}(E_0) = \mathbb{P}_{\pi_2}(\bigcup_{\alpha,\beta>0}\{N_{\alpha,\beta} = \infty\})$. Therefore there is $\alpha,\beta > 0$ such that $\mathbb{P}_{\pi_2}(\{N_{\alpha,\beta} = \infty\}) > 0$ and so $\mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] = \infty$.

Finally, we have by Proposition 6.3.5 that

$$\sup_{\alpha,\beta}\mathbb{E}_{\pi_2}\Big[N_{\alpha,\beta} \mid E_\ell\Big] \leq \mathbb{E}_{\pi_2}\Big[N_\perp \mid E_\ell\Big]$$

since by lemma 6.4.3, $L_n = 0$ if and only if we visit a bottom SCC C in the Markov Chain $\mathcal{D}$ such that $(\emptyset, q) \in C$ for some $q \in Q$. The conditional expected time of visiting a state in a Markov chain is finite. This follows directly from the main result of [43]. $\square$

7 Likelihood Exponents

In Section 6.4 we showed that computing the distribution over the possible values of $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ tends to be computationally difficult. In this chapter we focus on the problem of computing Λ. We show that any member of Λ can be expressed as the difference of two Lyapunov exponents. In general, computing a likelihood exponent is at least as hard as computing a Lyapunov exponent. Approximating or even computing Lyapunov exponents is hard, but there are practical approximation algorithms using convex optimisation [29, 30].

In Section 7.1 we study *deterministic* HMMs and show that this subclass leads to tractable problems. In Section 7.2 we introduce Lyapunov systems. We use Lyapunov systems to represent Lyapunov exponents and show that from a HMM one can easily compute a set of Lyapunov systems such that every possible likelihood exponent is the difference of two exponents represented by these systems. Here, we also show that the likelihood exponent exists almost surely and takes at most $|Q|^2 + 1$ values.

7.1 Deterministic HMMs

A HMM (Q, Σ, M) is *deterministic* if, for all $a \in \Sigma$, all rows of $M(a)$ contain at most one non-zero entry. Thus, for all $q \in Q$ and $w \in \Sigma^*$, we have $|\mathrm{supp}(e_q M(w))| \leq 1$.

A useful observation is that the numerator support Markov chain $\mathcal{D} = (2^Q \times Q, T)$, which was defined in Section 6.4 and can be exponential in general, has only quadratic size in the deterministic case if we restrict it to the part that is reachable from initial Dirac distributions.

Example 7.1.1. Consider the deterministic HMM (Q, Σ, M) in Figure 7.1(a). Let

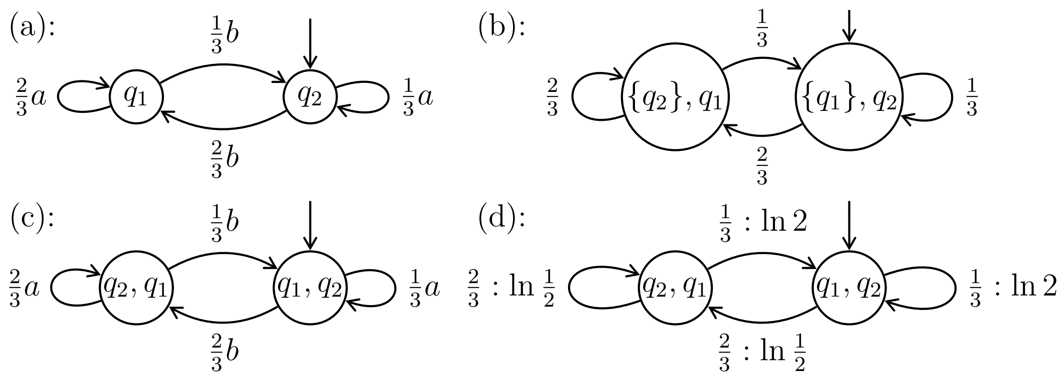


Figure 7.1: Cross-product constructions for a deterministic HMM.

$\pi_1 = e_{q_1}$ and $\pi_2 = e_{q_2}$ (the latter is indicated by an arrow pointing to q_2). Then the relevant (i.e., reachable from $(\{q_1\}, q_2)$) part of $\mathcal{D}$ is shown in Figure 7.1(b). Let us add back the observations that gave rise to the transitions in $\mathcal{D}$, and for simplicity drop the set brackets in the left component of states. We obtain the HMM in Figure 7.1(c). With this HMM we may keep track of the exact likelihood ratio. For example, suppose that the word aba is emitted, so that $L_3 = \frac{\|e_{q_1} M(aba)\|}{\|e_{q_2} M(aba)\|} = \frac{1}{2}$ and $\mathrm{supp}(e_{q_1} M(aba)) = \{q_2\}$ and $\mathrm{supp}(e_{q_2} M(aba)) = \{q_1\}$. Suppose the next letter is b (which is the case with probability $\frac{1}{3}$). Then L_4 arises from L_3 by multiplying with $\frac{M_{q_2,q_1}(b)}{M_{q_1,q_2}(b)} = 2$, and the supports are switched again. In terms of log-likelihoods, we have $\ln L_4 = \ln L_3 + \ln 2$. This motivates the Markov chain shown in Figure 7.1(d), where the transitions outgoing from a state (r_1, r_2) are labelled by the log-likelihood ratio of their corresponding probabilities in the HMM.

The Markov chain has stationary distribution $(\frac{2}{3}, \frac{1}{3})$. By the strong ergodic theorem for Markov chains, we obtain (the irrational number)

$$\lim_{n\to\infty} \frac{1}{n} \ln L_n \;=\; \frac{2}{3}\left(\frac{2}{3}\ln\frac{1}{2} + \frac{1}{3}\ln 2\right) + \frac{1}{3}\left(\frac{1}{3}\ln 2 + \frac{2}{3}\ln\frac{1}{2}\right) \;=\; \frac{1}{3}\ln 2 + \frac{2}{3}\ln\frac{1}{2} = -\frac{1}{3}\ln 2\,.$$

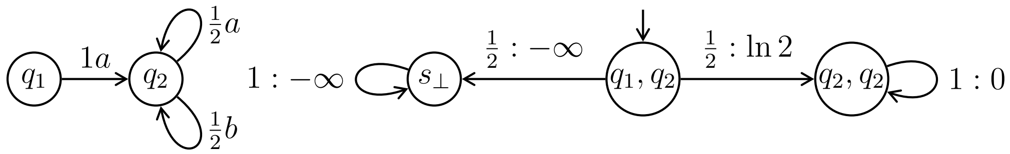
In general there may again be several likelihood exponents, including $-\infty$ and 0. For the rest of the section, let $\mathcal{F} = (Q, \Sigma, M)$ be a deterministic HMM. Motivated by

Example 7.1.1, define a HMM $\mathcal{A} = ((Q \times Q) \cup s_\perp, \hat{\Sigma}, \hat{M})$, where $s_\perp$ is a fresh state, and

$$\begin{aligned}
\hat{\Sigma} \;&:=\; \left\{ \ln \frac{M(a)_{q_1,r_1}}{M(a)_{q_2,r_2}} \in [-\infty, \infty) \;\middle|\; a \in \Sigma,\ q_1, r_1, q_2, r_2 \in Q,\ M(a)_{q_2,r_2} \neq 0 \right\} \cup \{-\infty\} \\
\hat{M}(\hat{a})_{(q_1,q_2),(r_1,r_2)} \;&:=\; \sum \left\{ M(a)_{q_2,r_2} \;\middle|\; a \in \Sigma : \hat{a} = \ln \frac{M(a)_{q_1,r_1}}{M(a)_{q_2,r_2}} \right\} \quad \text{for } \hat{a} \neq -\infty \\
\hat{M}(-\infty)_{(q_1,q_2),s_\perp} \;&:=\; \sum \left\{ M(a)_{q_2,r_2} \;\middle|\; a \in \Sigma,\ r_2 \in Q : \textstyle\sum_{r_1 \in Q} M(a)_{q_1,r_1} = 0 \right\} \\
\hat{M}(-\infty)_{s_\perp,s_\perp} \;&:=\; 1\,.
\end{aligned}$$

Note that the embedded Markov chain of $\mathcal{A}$ is similar to the Markov chain $\mathcal{D}$ from Lemma 6.4.3: states $(\{q_1\}, q_2)$ in $\mathcal{D}$ are called (q_1, q_2) in $\mathcal{A}$, the states $(\emptyset, q)$ in $\mathcal{D}$ are subsumed by the state $s_\perp$ of $\mathcal{A}$, and the states (S, q) in $\mathcal{D}$ with $|S| > 1$ are not represented in $\mathcal{A}$. The observations in $\hat{\Sigma} \subseteq [-\infty, \infty)$ track the log-likelihood ratio.

Example 7.1.2. Consider the HMM $\mathcal{F}$ on the left, with initial distributions $\pi_1 = e_{q_1}$ and $\pi_2 = e_{q_2}$. The part of $\mathcal{A}$ reachable from (q_1, q_2) is shown on the right:



Here we have $\Lambda_{\pi_1,\pi_2} = \{-\infty, 0\}$ with $\mathbb{P}_{\pi_2}(E_{-\infty}) = \mathbb{P}_{\pi_2}(E_0) = \frac{1}{2}$.

Denote by $\bar{\mathcal{A}}$ the embedded Markov chain of $\mathcal{A}$. Let $C \subseteq Q \times Q$ be a non-$\{s_\perp\}$ bottom SCC of $\bar{\mathcal{A}}$. Let $\mu \in [0,1]^C$ denote the stationary distribution of the restriction of $\bar{\mathcal{A}}$ on C. Define the vector $\nu \in \mathbb{R}^C$ of average observations by $\nu_{(r_1,r_2)} := \sum_{\hat{a} \in \hat{\Sigma}} \|e_{(r_1,r_2)} \hat{M}(\hat{a})\| \cdot \hat{a}$. By the strong ergodic theorem for Markov chains, the *average observation* in C equals $\mu\nu^\top =: \ell(C)$. Extend this definition by $\ell(\{s_\perp\}) := -\infty$. Then we have the following lemma.

Lemma 7.1.3. *Let $\pi_1 = e_{q_1}$ and $\pi_2 = e_{q_2}$ be initial distributions. For the Markov chain $\bar{\mathcal{A}}$ define $\iota := e_{(q_1,q_2)}$. We have*

$$\mathbb{P}_{\pi_2}(E_\ell) = \mathbb{P}_\iota(\{\textit{visit bottom SCC } C \textit{ with } \ell(C) = \ell\}).$$

The proof is essentially the same as in the third point of Lemma 6.4.3 which gives us the following result.

Theorem 7.1.4. *Given a deterministic HMM (Q, Σ, M) with initial Dirac distributions π_1, π_2, one can compute in polynomial time*

1. Λ_{π_1,π_2} *as a set of expressions of the form* $\sum_i x_i \ln y_i$ *where* $x_i, y_i \in \mathbb{Q}$, *and*

2. $\Pr_{\pi_2}(E_\ell)$ *for each such* $\ell \in \Lambda_{\pi_1,\pi_2}$.

Proof. In a Markov chain, one can compute the stationary distribution and hitting probabilities in polynomial time by solving a linear system of equations. Thus, the numbers $\ell(C)$ defined before Lemma 7.1.3 can be computed in polynomial time. Both parts of the theorem follow then from Lemma 7.1.3. A slight complication is that for part 2, for an $\ell = \sum_i x_i \ln y_i \in \Lambda_{\pi_1,\pi_2}$, in order to compute $\mathbb{P}_{\pi_2}(E_\ell)$ we have to sum the hitting probabilities for all C with $\ell = \ell(C)$. To select those C we have to compare numbers of the form $\sum_i x_i \ln y_i$ where $x_i, y_i \in \mathbb{Q}$. This is equivalent to checking multiplicative relations among rationals. The test takes polynomial time as shown in [44]. $\square$

7.2 Representing Likelihood Exponents

In the following we show that one can efficiently represent likelihood exponents in terms of *Lyapunov exponents*. The definition of Lyapunov exponents is based on the following definition.

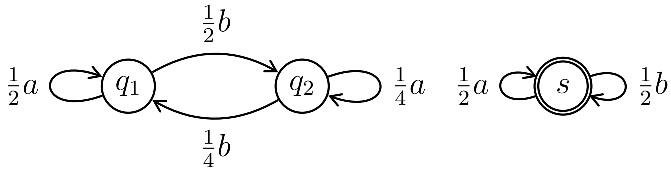
Definition 7.2.1. A *matrix system* is a triple $\mathcal{M} = (Q, \Sigma, M)$ where Q is a finite set of states, Σ is a finite set of observations, and $M : \Sigma \to \mathbb{R}_{\geq 0}^{Q \times Q}$ specifies the transitions. (Note that a HMM is a matrix system.) A *Lyapunov system* is a pair $\mathcal{S} = (\mathcal{M}, \rho)$ where $\mathcal{M} = (Q, \Sigma, M)$ is a matrix system and $\rho \in (0,1]^\Sigma$ is a probability distribution with full support, such that the directed graph (Q, E) with $E = \{(q, r) \mid \sum_{a \in \Sigma} M_{q,r}(a) > 0\}$ is strongly connected.

We can identify the probability distribution ρ from this definition with the single-state HMM $(\{s\}, \Sigma, M_\rho)$ where $M_\rho(a)_{s,s} = \rho(a)$ for all $a \in \Sigma$. In this way, ρ produces a random infinite word from Σ^ω. We will write $\mathbb{P}_\rho$ for the associated probability measure.

Example 7.2.1. Let $\mathcal{M} = (Q, \Sigma, M)$ be a matrix system where $Q = \{q_1, q_2\}$, $\Sigma = \{a, b\}$ and

$$M(a) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}, \quad M(b) = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{4} & 0 \end{pmatrix}.$$

Further, let $\rho = (\frac{1}{2}, \frac{1}{2})$. Then, $(\mathcal{M}, \rho)$ is a Lyapunov system. We can represent a Lyapunov system similarly to the way we represent a HMM. We use the double ringed state for the single-state HMM identified with ρ.

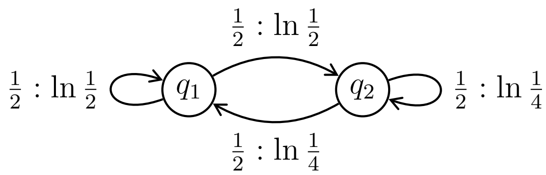


The following lemma is Theorem 1 from [45].

Lemma 7.2.2 ([45])**.** *Let $((Q, \Sigma, M), \rho)$ be a Lyapunov system. Then there is $\lambda \in \mathbb{R}$ such that, for all $\pi \in [0, \infty)^Q$, $\mathbb{P}_\rho$-a.s., either $\pi M(w_n) = \overline{0}$ for some $n \in \mathbb{N}$ or the limit $\lim_{n\to\infty} \frac{1}{n} \ln \|\pi M(w_n)\|$ exists and equals λ.*

For a Lyapunov system $\mathcal{S}$ we call $\lambda(\mathcal{S}) = \lambda$ from the lemma the *Lyapunov exponent* defined by $\mathcal{S}$. In the case that $\mathbb{P}_{\pi_2}$-a.s. for all $\pi \in [0,1]^Q$ there is $n \in \mathbb{N}$ such that $M(w_n) = \overline{0}$, we say $\lambda(\mathcal{S}) = -\infty$.

Example 7.2.3. We continue with Example 7.2.1. Let $w \in \Sigma^\omega$ and $n \in \mathbb{N}$. Like with Markov chain (d) from Example 7.1.1 we construct a Markov chain to keep track of the difference $\ln \|e_{q_1} M(w_n)\| - \ln \|e_{q_1} M(w_{n-1})\|$ which we label on each transition.



The Markov chain has stationary distribution $(\frac{1}{2}, \frac{1}{2})$. By the strong ergodic theorem for Markov chains we have for $i = 1, 2$:

$$\lim_{n\to\infty} \ln \frac{1}{n} \|e_{q_i} M(w_n)\| = \frac{1}{2} \ln \left(\frac{1}{2}\right) + \frac{1}{2} \ln \left(\frac{1}{4}\right) = -\frac{3}{2} \ln 2.$$

While we cannot in general compute Lyapunov exponents this way, we can represent them as the difference of two Lyapunov exponents corresponding to easily computable Lyapunov systems. In Section 7.2.4 we prove the following theorem, which implies Theorem 6.2.3.

Theorem 7.2.4. *Given a HMM $\mathcal{F} = (Q, \Sigma, M)$ we can compute in polynomial time $2K \leq 2|Q|^2$ Lyapunov systems $\mathcal{S}_1^1, \mathcal{S}_1^2, \mathcal{S}_2^1, \mathcal{S}_2^2, \ldots, \mathcal{S}_K^1, \mathcal{S}_K^2$ such that for any initial distributions π_1, π_2 the limit $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ exists $\mathbb{P}_{\pi_2}$-a.s. and lies in*

$$\Lambda \subseteq \{-\infty\} \cup \{\lambda(\mathcal{S}_1^1) - \lambda(\mathcal{S}_1^2), \ldots, \lambda(\mathcal{S}_K^1) - \lambda(\mathcal{S}_K^2)\}\,.$$

In particular, $\mathcal{F}$ has at most $|Q|^2 + 1$ likelihood exponents.

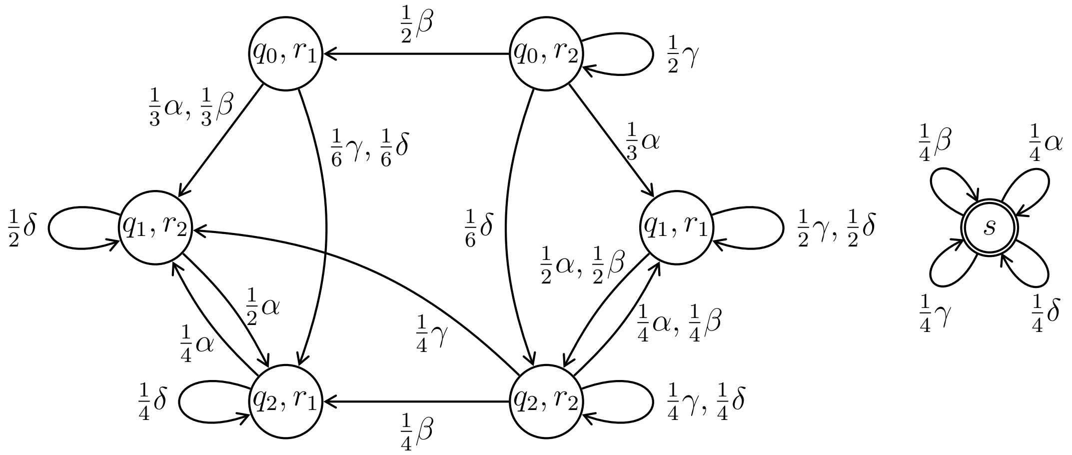
The proof of Theorem 7.2.4 proceeds by several generalisations of Lemma 7.2.2.

1. In Section 7.2.1 we relax the assumption that Q is strongly connected by introducing *reducible Lyapunov systems*. A characterisation of the possible Lyapunov exponents in this case is proven in Lemma 7.2.9.
2. In Section 7.2.2 we relax the assumption that the observations are produced independently by the distribution ρ and assume the letters are produced by a strongly connected HMM by introducing *generalized Lyapunov systems*. A characterisation of the possible Lyapunov exponents in this case is stated in the first point of Lemma 7.2.12. In Section 7.2.3 we prove Lemma 7.2.12.
3. In Section 7.2.4, we relax all assumptions. In Lemma 7.2.16 we show that a Likelihood exponent can be written as the difference of two Lyapunov exponents. These Lyapunov exponents correspond to particular generalized Lyapunov systems which can be transformed into Lyapunov systems by Lemma 7.2.12. By constructing a suitable reducible Lyapunov system we can characterise the possible values of the Likelihood exponent by Lemma 7.2.9. Theorem 7.2.4 follows quickly from Lemma 7.2.16.

7.2.1 Reducible Lyapunov Systems

A *reducible Lyapunov system* is a Lyapupov system $((Q, \Sigma, M), \rho)$ except that the embedded graph $G = (Q, E)$ with $E = \{(q, r) \mid \exists\, a \in \Sigma : M_{q,r}(a) > 0\}$ is not necessarily strongly connected. In this section we introduce new notation. For row vectors $\pi_1 \in [0, \infty)^{Q_1}$ and $\pi_2 \in [0, 1]^{Q_2}$ we write $\pi_1 \times \pi_2 \in [0, \infty]^{Q_1 \times Q_2}$ for the vector $(\pi_1 \times \pi_2)_{(i,j)} := (\pi_1)_i (\pi_2)_j$. In addition, for a function of the form $\Phi : \Sigma \to \mathbb{R}^{Q\times Q}$ and $P \subseteq Q$ we write $\Phi_{|P} : \Sigma \to \mathbb{R}^{P\times P}$ for the function with $\Phi_{|P}(a)(q, r) = \Phi(a)(q, r)$ for all $a \in \Sigma$ and $q, r \in P$; i.e., $\Phi_{|P}(a)$ denotes the principal submatrix obtained from $\Phi(a)$ by restricting it to the rows and columns indexed by P.

Example 7.2.5. We present the following reducible Lyapunov system. Let $Q = \{q_0, q_1, q_2\} \times \{r_1, r_2\}$, $\Delta = \{\alpha, \beta, \gamma, \delta\}$ and $\rho = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Then $((Q, \Delta, M), \rho)$ is a reducible Lyapunov system where the transitions in M are shown below.



Since there is no path in the embedded graph from (q_1, r_2) to (q_0, r_1), the embedded graph is not strongly connected and so $((Q, \Delta, M), \rho)$ is not a Lyapunov system.

We use the following classical result from [46].

Theorem 7.2.6 (Fürstenberg–Kesten theorem)**.** *Let $((Q, \Sigma, M), \mu)$ be a reducible Lyapunov system. Then, there exists $\lambda \in [-\infty, \infty)$ such that $\mathbb{P}_\rho$-a.s. we have*

$$\lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}M(w_n)\| = \lambda. \tag{7.1}$$

Example 7.2.7. We simulate the value of $\frac{1}{n} \ln \|\bar{1}M(w_n)\|$ as $n \to \infty$ in Figure 7.2. We see that the value of λ from Theorem 7.2.6 is approximately $\lambda \approx -1$.

The following lemma is Theorem 1.1 from [47].

Lemma 7.2.8. *Let Σ be a finite set of observations and $\rho \in [0,1]^\Sigma$ be a distribution on the set of observations. Let $((Q_1, \Sigma, M_1), \rho)$ and $((Q_2, \Sigma, M_2), \rho)$ be reducible Lyapunov systems. By Theorem 7.2.6 there exists $\lambda_1, \lambda_2 \in [-\infty, \infty)$ such that for $i = 1, 2$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}M_i(w_n)\| = \lambda_i.$$

Further, let $M_{12} : \Sigma \to [0,1]^{Q_1 \times Q_2}$ specify transition from Q_1 to Q_2 and define $\hat{M} : \Sigma \to [0,1]^{(Q_1 \cup Q_2) \times (Q_1 \cup Q_2)}$ block-wise as

$$\hat{M}(a) = \begin{pmatrix} M_1(a) & M_{12}(a) \\ 0 & M_2(a) \end{pmatrix}.$$

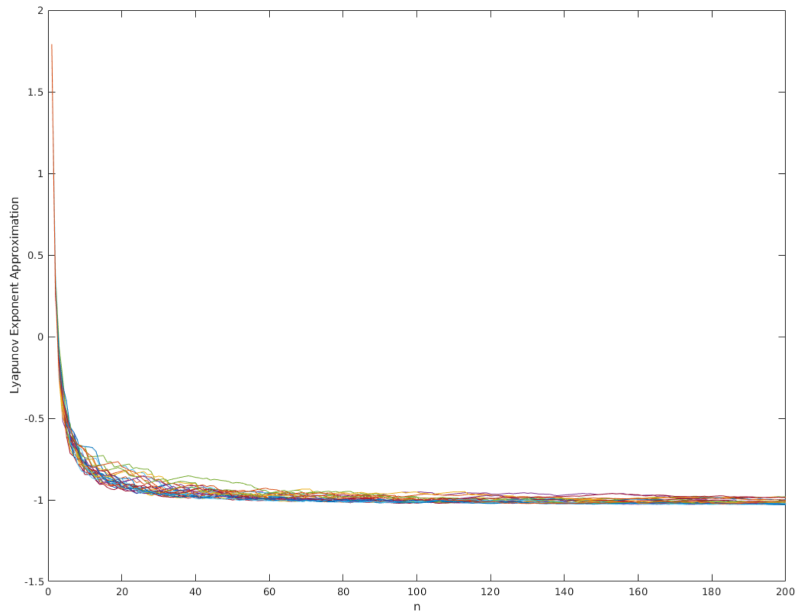


Figure 7.2: Approximating the Lyapunov exponent of Figure 7.2 with 30 samples and $n \leq 200$.

It follows $((Q_1 \cup Q_2, \Sigma, \hat{M}), \rho)$ *is a reducible Lyapunov system and*

$$\lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}\hat{M}(w_n)\| = \max\{\lambda_1, \lambda_2\}.$$

Recall that for sets $U, V \subseteq Q$ we write $U \to_G V$ if there is $u \in U$ and $v \in V$ such that v is reachable from u in G. The following is a generalisation of Lemma 7.2.8 when the induced graph has more than two strongly connected components.

Lemma 7.2.9. *Let* $((Q, \Sigma, M), \rho)$ *be a reducible Lyapunov system and suppose* (Q, E) *has SCCs* $C_1, \ldots, C_K$*. Let* $\mathbf{0}$ *be the zero matrix and let*

$$\mathcal{Z} = \{C \in \{C_1, \ldots, C_K\} \mid \exists u \in \Sigma^* \text{ s.t. } M_{|C}(u) = \mathbf{0}\}.$$

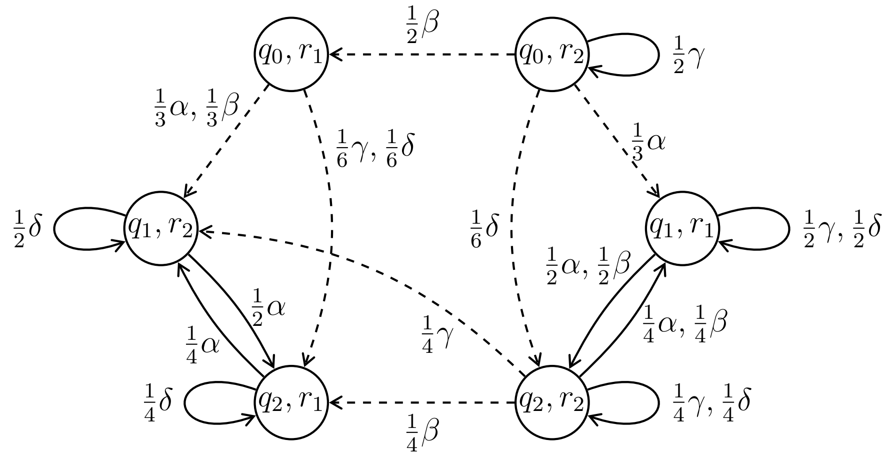
For each $k \in [K]$*,* $\mathcal{G}_k := ((C_k, \Sigma, M_{|C_k}), \rho)$ *is a Lyapunov system. We have that* $\mathbb{P}_\rho$*-almost surely,*

$$\lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}M(w_n)\| = \max\left(\left\{\lambda(\mathcal{G}_k) \mid C_k \notin \mathcal{Z}\right\} \cup \{-\infty\}\right) \tag{7.2}$$

where $\bar{1}$ *is the row vector all of whose entries are* 1*. Further, for any initial distribution* $\pi \in [0,1]^Q$*, we have that* $\mathbb{P}_\rho$*-almost surely,*

$$\lim_{n\to\infty} \frac{1}{n} \ln \|\pi M(w_n)\| \in \left\{\lambda(\mathcal{G}_k) \mid \mathit{supp}(\pi) \to_G C_k, C_k \notin \mathcal{Z}\right\} \cup \{-\infty\}. \tag{7.3}$$

Example 7.2.10. We show the matrix system from Example 7.2.5 with any transitions between SCCs of the embedded graph displayed with dashed lines.



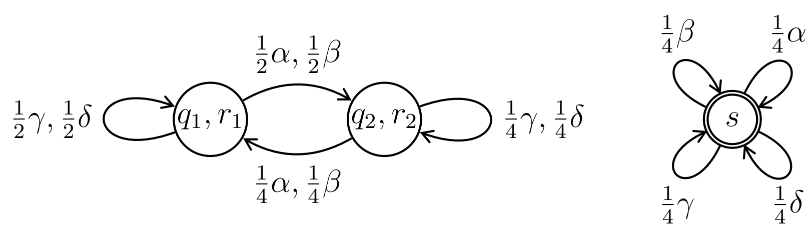
The SCCs of the induced graph are $\{(q_0, r_2)\}, \{(q_0, r_1)\}, \{(q_1, r_1), (q_2, r_2)\}$ and $\{(q_1, r_2), (q_2, r_1)\}$. We have that

$$\begin{aligned} M_{|\{(q_0,r_2)\}}(\alpha) &= 0, \\ M_{|\{(q_0,r_1)\}}(\alpha) &= 0 \quad \text{and} \\ M_{|\{(q_1,r_1),(q_2,r_2)\}}(\beta) &= 0. \end{aligned}$$

On the other hand, $\|\bar{1}M_{|\{(q_1,r_1),(q_2,r_2)\}}(u)\| > 0$ for all $u \in \Delta^*$. Therefore we compute the set $\mathcal{Z}$ from Lemma 7.2.9 as $\mathcal{Z} = \{\{(q_0, r_2)\}, \{(q_0, r_1)\}, \{(q_1, r_2), (q_2, r_1)\}\}$. By (7.2) we have that $\mathbb{P}_\rho$-almost surely,

$$\lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}M(w_n)\| = \lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}M_{|\{(q_1,r_1),(q_2,r_2)\}}(w_n)\|.$$

We show the Lyapunov system $((\{(q_1, r_1), (q_2, r_2)\}, \Delta, M_{\{(q_1,r_1),(q_2,r_2)\}}), \rho)$ below.



Let $\mathcal{S}' = ((Q', \Sigma', M'), \rho)$ be the Lyapunov system from Example 7.2.1 where $\Sigma = \{a', b'\}$ and $\rho = (\frac{1}{2}, \frac{1}{2})$. We have $M'(b') = M_{\{(q_1,r_1),(q_2,r_2)\}}(\alpha) = M_{\{(q_1,r_1),(q_2,r_2)\}}(\beta)$

and $M'(a') = M_{\{(q_1,r_1),(q_2,r_2)\}}(\gamma) = M_{\{(q_1,r_1),(q_2,r_2)\}}(\delta)$. Therefore, $\mathbb{P}_\rho$-almost surely,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\ln\|\overline{1}M(w_n)\| &= \lim_{n\to\infty}\frac{1}{n}\ln\|\overline{1}M_{|\{(q_1,r_1),(q_2,r_2)\}}(w_n)\| \\ &= \lambda(\mathcal{S}') \\ &= -\frac{3}{2}\ln 2 \\ &\approx -1.\end{aligned}$$

This coincides with our estimate from Example 7.2.7.

With the embedded graph G from Lemma 7.2.9, we have that $(q_0, r_2) \to_G \{(q_1,r_1),(q_2,r_2)\}$. Hence, $\mathbb{P}_\rho$-almost surely,

$$\lim_{n\to\infty}\frac{1}{n}\ln\|e_{(q_0,r_2)}M(w_n)\| \in \{-\frac{3}{2}\ln 2, -\infty\}.$$

The word $\beta\beta\beta \in \Delta^3$ is a mortal word for M started from state (q_0, r_2) and so the Lyapunov exponent $-\infty$ is realised with positive probability.

Proof of Lemma 7.2.9. We first prove (7.2). By Theorem 7.2.6 for $i \in [N]$ there exists $\lambda_i \in [-\infty, \infty)$ such that $\mathbb{P}_\rho$-a.s.,

$$\lim_{n\to\infty}\frac{1}{n}\ln\|\overline{1}M_{|C_i}(w_n)\| = \lambda_i.$$

In the case $\lambda_i = -\infty$ then by Lemma 7.2.2 we have that $\mathbb{P}_\rho$-a.s. there is a word $u \in \Sigma^*$ such that $\overline{1}M_{|C_i}(u) = \overline{0}$ hence $C_i \in \mathcal{Z}$.

In the case $\lambda_i \in \mathbb{R}$ then by Lemma 7.2.2 $\lambda_i = \lambda(\mathcal{G}_i)$. Also $\{w \in \Sigma^\omega \mid \exists n \in \mathbb{N} : \overline{1}M_{|C_i}(w_n) = \overline{0}\}$ is $\mathbb{P}_\rho$-null set and therefore empty because ρ has full support and therefore all finite words have positive probability of being produced. It follows that $C_i \notin \mathcal{Z}$.

We proceed by induction. In the case $K = 1$, the cases described previously immediately imply that (7.2) holds. Now we assume the lemma holds for some $K = m \in \mathbb{N}$. We prove this implies the theorem holds for $K = m + 1$.

Let $((Q, \Sigma, M), \rho)$ be a reducible Lyapunov system such that (Q, E) has strongly connected components $C_1, \dots, C_{m+1}$. We have that $((C_1 \cup \dots \cup C_m, \Sigma, M_{|C_1\cup\dots\cup C_m}), \rho)$ is a reducible Lyapunov system so by the induction hypothesis, $\mathbb{P}_\rho$-a.s. we have

$$\lim_{n\to\infty}\frac{1}{n}\ln\|\overline{1}M_{|C_1\cup\dots\cup C_m}(w_n)\| = \max\Big(\{\lambda(\mathcal{G}_i) \mid i \in [m], C_i \notin \mathcal{Z}\} \cup \{-\infty\}\Big).$$

In the case that $\lambda_{m+1} \in \mathbb{R}$ we have $\lambda(\mathcal{G}_{m+1}) = \lambda_{m+1}$ and $C_{m+1} \notin \mathcal{Z}$. Therefore, $\lambda_{m+1} \in \{\lambda(\mathcal{G}_i) \mid i \in [m+1], C_i \notin \mathcal{Z}\}$.

In the case $\lambda_{m+1} = -\infty$, then $C_{m+1} \in \mathcal{Z}$ and clearly $\{\lambda(\mathcal{G}_i) \mid i \in [m], C_i \notin \mathcal{Z}\} = \{\lambda(\mathcal{G}_i) \mid i \in [m+1], C_i \notin \mathcal{Z}\}$. So by Lemma 7.2.8 we have

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}M(w_n)\| &= \max\Big\{\lambda_{m+1}, \max\Big(\{\lambda(\mathcal{G}_i) \mid i \in [m], C_i \notin \mathcal{Z}\} \cup \{-\infty\}\Big)\Big\} \\ &= \max\Big(\{\lambda(\mathcal{G}_i) \mid i \in [m+1], C_i \notin \mathcal{Z}\} \cup \{-\infty\}\Big).\end{aligned}$$

We now prove (7.3). Like in Section 6.4, for $S \subseteq Q$ and $a \in \Sigma$, define $\delta'(S,a) = \{q' \in Q \mid \exists\, q \in S : M(a)_{q,q'} > 0\}$. Then we define the Markov chain $\mathcal{D}' := (2^Q, T')$ where

$$T'_{S,S'} := \sum_{\delta'(S,a)=S'} \rho(a).$$

Let $\mathcal{D}$ be the set of bottom SCCs that are reachable in $\mathcal{D}'$ from the initial state $\operatorname{supp}(\pi)$. We fix $D \in \mathcal{D}$. We may order the $C_1 \cdots C_K$ such that for some $k \in [K]$

$$S_D := \{q \in S \mid S \in D\} \cap C_i \neq \emptyset$$

for all $i \in [k]$. It is clear that $S_D \subseteq C_1 \cup \cdots \cup C_k$. We show the reverse inclusion. Let $i \in [k]$ then there is $q_0 \in C_i \cap S_D$. For any $q' \in C_i$ there is a word $a_1 q_1 \cdots a_m q_m \in (\Sigma Q)^*$ such that $q_m = q'$ and $M_{q_{i-1},q_i}(a_i) > 0$ for all $i \in [m]$. It follows that $q' \in S_D$ and therefore $S_D = C_1 \cup \cdots \cup C_k$.

We have that $\mathbb{P}_\rho$-a.s. $\lim_{n\to\infty} \frac{1}{n} \ln \|\bar{1}M_{|S_D}(w_n)\|$ exists and equals

$$\lambda_D := \max\Big(\{\lambda(\mathcal{G}_i) \mid i \in [k], C_i \notin \mathcal{Z}\} \cup \{-\infty\}\Big)$$

by (7.2). In particular there is $q \in S_D$ such that

$$\alpha := \mathbb{P}_\rho\Big(\lim_{n\to\infty} \frac{1}{n} \ln \|e_q M(w_n)\| = \lambda_D\Big) = \mathbb{P}_\rho\Big(\lim_{n\to\infty} \frac{1}{n} \ln \|e_q M_{|S_D}(w_n)\| = \lambda_D\Big) > 0.$$

For all $S_0 \in D$ there is a sequence $a_1 S_1 \cdots a_n S_n \in (\Sigma D)^*$ such that $q \in S_n$ and $S_i = \delta(S_{i-1}, a_i)$ for all $i = 1, \ldots, n$. We have that $\beta_{S_0} := \prod_{i=1}^n \rho(a_i) > 0$. Write $\beta = \min\{\beta_{S_0} \mid S_0 \in D\}$. Let $\pi' \in [0,1]^Q$ be such that $\operatorname{supp}(\pi') \in D$. Then for any

$u \in \Sigma^*$ we have

$$\begin{aligned}
&\mathbb{P}_\rho(\lim_{n\to\infty} \frac{1}{n} \ln \|\pi' M(w_n)\| = \lambda_D \mid u\Sigma^\omega) \\
&= \mathbb{P}_\rho(\lim_{n\to\infty} \frac{1}{n} \ln \|\pi' M(w_n)\| = \lambda_D \mid u\Sigma^\omega) \\
&\geq \beta\mathbb{P}_\rho(\lim_{n\to\infty} \frac{1}{n} \ln \|\pi' M(uw_n)\| = \lambda_D) \\
&\geq \beta\mathbb{P}_\rho(\lim_{n\to\infty} \frac{1}{n} \ln \|e_q M(w_n)\| = \lambda_D) \\
&= \beta\alpha > 0.
\end{aligned}$$

Hence by Lévy's 0-1 law we have

$$\begin{aligned}
1 &= \mathbb{P}_\rho\Big(\lim_{n\to\infty} \frac{1}{n} \ln \|\pi' M(w_n)\| = \lambda_D\Big) \\
&= \mathbb{P}_\rho\Big(\lim_{n\to\infty} \frac{1}{n} \ln \|\pi' M(w_n)\| \in \{\lambda(\mathcal{G}_i) \mid \mathrm{supp}\,(\pi) \to_G C_i, C_i \notin \mathcal{Z}\} \cup \{-\infty\}\Big)
\end{aligned}$$

where the last equality follows because $D \in \mathcal{D}$ is reachable from $\mathrm{supp}\,(\pi)$ in $\mathcal{D}'$ which implies that for any $C_i \subseteq S_D$ we have $\mathrm{supp}\,(\pi) \to_G C_i$.

We now relax the dependence on D. For any $D \in \mathcal{D}$, let $F_D = \{w \in \Sigma^\omega \mid \exists n \in \mathbb{N} : \mathrm{supp}\,(\pi)M(w_n)) \in D\}$. Then,

$$\mathbb{P}_\rho\Big(\bigcup_{D\in\mathcal{D}} F_D\Big) = 1.$$

Write $\Theta = \{\lambda(\mathcal{G}_i) \mid \mathrm{supp}\,(\pi) \to_G C_i, C_i \notin \mathcal{Z}\} \cup \{-\infty\}$ then $\mathbb{P}_\rho$-a.s. we have

$$\begin{aligned}
\mathbb{P}_\rho(\lim_{n\to\infty} \frac{1}{n} \ln \|\pi M(w_n)\| \in \Theta) &= \sum_{D\in\mathcal{D}} \mathbb{P}_\rho(\{\lim_{n\to\infty} \frac{1}{n} \ln \|\pi M(w_n)\| \in \Theta\} \mid F_D)\mathbb{P}_\rho(F_D) \\
&= \sum_{D\in\mathcal{D}} \mathbb{P}_\rho(F_D) = 1.
\end{aligned}$$

□

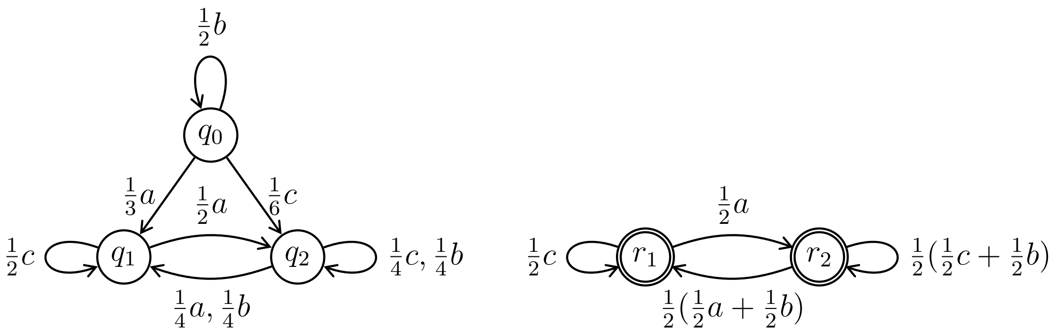
7.2.2 Generalised Lyapunov Systems

In this section we define *generalized Lyapunov systems.*

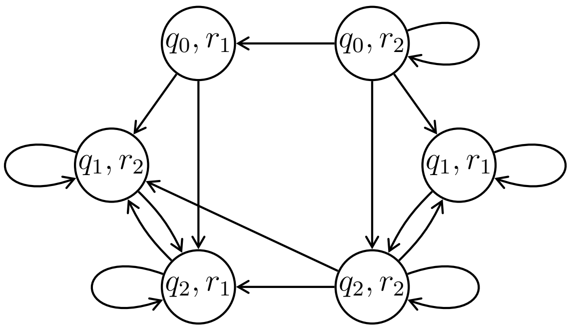
First, let $\mathcal{M}_1 = (Q_1, \Sigma, M_1)$ and $\mathcal{M}_2 = (Q_2, \Sigma, M_2)$ be matrix systems with finite Q_1, Q_2, Σ and transitions $M_i : \Sigma \to \mathbb{R}_{\geq 0}^{Q_i \times Q_i}$ for $i = 1, 2$. Like with HMMs, we define the joint graph. Let $G_{\mathcal{M}_1,\mathcal{M}_2} = (Q_1 \times Q_2, E)$ such that there is an edge from (q_1, q_2) to (r_1, r_2) if there is $a \in \Sigma$ with $M_1(a)_{q_1,r_1} > 0$ and $M_2(a)_{q_2,r_2} > 0$.

A *generalized Lyapunov system* is a triple $\mathcal{S} = (\mathcal{M}, \mathcal{F}, C)$ where $\mathcal{M} = (Q_1, \Sigma, M_1)$ is a matrix system and $\mathcal{F} = (Q_2, \Sigma, M_2)$ is a strongly connected HMM and $C \subseteq Q_1 \times Q_2$ is a bottom SCC of $G_{\mathcal{M},\mathcal{F}}$.

Example 7.2.11. Recall the HMM $\mathcal{F} = (Q, \Sigma, M)$ from Chapter 4 in Example 4.1.1. We have $\Sigma = \{a, b, c\}$. Consider the diagram below where on the left we give an example matrix system $\mathcal{M} = (Q_1, \Sigma, M_1)$ and on the right we give the HMM $\mathcal{F}$. The HMM $\mathcal{F}$ "produces" the letters and so similarly to how we draw Lyapunov system diagrams, we give the states of $\mathcal{F}$ a double ring.



We also show the joint graph $G_{\mathcal{M},\mathcal{F}}$ below which has a single bottom SCC $C = \{(q_1, r_2), (q_2, r_1)\}$.



Therefore, $(\mathcal{M}, \mathcal{F}, C)$ is a generalised Lyapunov system.

We show in the following subsection that given a generalized Lyapunov system, one can efficiently compute an "equivalent" Lyapunov system:

Lemma 7.2.12. *Let $\mathcal{S} = ((Q_1, \Sigma, M_1), (Q_2, \Sigma, M_2), C)$ be a generalized Lyapunov system.*

1. *There is $\lambda \in \mathbb{R}$, henceforth called $\lambda(\mathcal{S})$, such that, for all $\pi_1 \in [0, \infty)^{Q_1}$ and all probability distributions $\pi_2 \in [0, 1]^{Q_2}$ with supp$(\pi_1) \times$ supp$(\pi_2) \subseteq C$, we have $\mathbb{P}_{\pi_2}$-a.s. that either $\pi_1 M_1(w_n) = \overline{0}$ for some $n \in \mathbb{N}$ or the limit $\lim_{n\to\infty} \frac{1}{n} \ln \|\pi_1 M_1(w_n)\|$ exists and equals $\lambda(\mathcal{S})$.*

2. *One can compute in polynomial time a Lyapunov system $\mathcal{S}'$ such that $\lambda(\mathcal{S}) = \lambda(\mathcal{S}')$.*

7.2.3 Independent Production of Observations

From a HMM $\mathcal{F} = (Q, \Sigma, M)$, define $M'(a,r)_{q,r} := M(a)_{q,r}$ for all $a \in \Sigma$ and $q, r \in Q$. Then, $(Q, \Sigma \times Q, M')$ is a HMM which is similar to $\mathcal{F}$ but emits the next state, in addition to an observation from Σ. The following is the key technical lemma that along with Lemma 7.2.9 implies both points in Lemma 7.2.12.

Lemma 7.2.13. *Let $\mathcal{F} = (Q, \Sigma, M)$ be a HMM and define M' as in the main text such that $\mathcal{F}' = (Q, \Sigma Q, M')$ is a HMM. One can compute in polynomial time a finite set Δ, a probability distribution $\rho \in (0,1]^{\Delta}$ and for each $r_0 \in Q$ (a representation of) a mapping $\kappa_{r_0} : \Delta \to (\Sigma Q)$ with the following property.*

Extend κ_{r_0} to $(\Sigma Q)^+$ inductively. For each $c \in \Delta$ and $\overline{w} \in \Delta^+$ we let $\kappa_{r_0}(c\,\overline{w}) = (a_1 r_1)\kappa_{r_1}(\overline{w})$ where $a_1 r_1 = \kappa_{r_0}(c)$. Then, for each word $w \in (\Sigma Q)^+$ we have

$$\mathbb{P}_\rho(\kappa_{r_0}^{-1}(\{w\})\Delta^\omega) = \mathbb{P}_{e_{r_0}}(w(\Sigma Q)^\omega) \tag{7.4}$$

where $\mathbb{P}_{e_{r_0}}$ refers to the measure on words produced by $\mathcal{F}'$ with initial distribution e_{r_0}.

Proof. Let $\tau : [\Sigma Q] \to \Sigma Q$ be a bijection which we view as an arbitrary ordering on ΣQ and write $\tau(k) = a_k q_k$. For each $r \in Q$ we define a function $\overline{\kappa}_r : [0,1) \to \Sigma Q$

$$\overline{\kappa}_r(x) = a_K q_K \text{ when } \sum_{k=1}^{K-1} M(a_k)_{r,q_k} \le x < \sum_{k=1}^{K} M(a_k)_{r,q_k},$$

where for notational purposes, $\sum_{k=1}^{0} M(a_k)_{p,q_k} = 0$. For each $r \in Q$, since $\sum_{k=1}^{|Q\times\Sigma|} M(a_k)_{r,q_k} = 1$ it follows that $\overline{\kappa}_r$ is well defined on the interval $[0,1)$. Let Δ be the set of atomic elements of the finite σ-algebra $\sigma\{\overline{\kappa}_r^{-1}(\{aq\}) \mid r, q \in Q, a \in \Sigma\}$. The set Δ is a finite partition of $[0,1)$ and consists of intervals $[\alpha, \beta)$ where $\alpha, \beta \in [0,1] \cap \mathbb{Q}$.

One can compute in polynomial time the endpoints of all intervals $[\alpha, \beta) \in \Delta$. For any $a \in \Delta$ and $r \in Q$ the image $\overline{\kappa}_r(a)$ contains exactly one element. Therefore, for each $r \in Q$ we may define a function $\kappa_r : \Delta \to \Sigma Q$ such that $\kappa_r([\alpha, \beta)) = \overline{\kappa}_r(x)$ for all $x \in [\alpha, \beta)$. We define the probability distribution ρ on Δ by $\rho_{[\alpha,\beta)} = \beta - \alpha$ for all $[\alpha, \beta) \in \Delta$. The distribution ρ is computable in polynomial time.

Let $r_0 \in Q$. We may extend the mapping κ_{r_0} to a word $c\,\overline{w} \in \Delta^+$ inductively by letting $\kappa_{r_0}(c\,\overline{w}) = (a_1 r_1)\kappa_{r_1}(\overline{w})$ where $(a_1 r_1) = \kappa_{r_0}(c)$. We now prove (7.4) by induction on the length of the word. Let $a_1 r_1 \in \Sigma Q$ then by the definition of κ_{r_0},

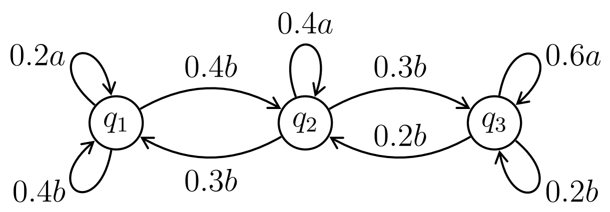
$$\mathbb{P}_{e_{r_0}}(a_1 r_1 (\Sigma Q)^\omega) = M(a_1)_{r_0,r_1} = \rho\Big(\kappa_{r_0}^{-1}(\{a_1 r_1\})\Big) = \mathbb{P}_\rho(\kappa_{r_0}^{-1}(\{a_1 r_1\})\Delta^\omega).$$

Now assume Equation (7.4) holds for all words of length $n-1 \in \mathbb{N}$. Then for a word $a_1 r_1 w \in (\Sigma Q)^n$

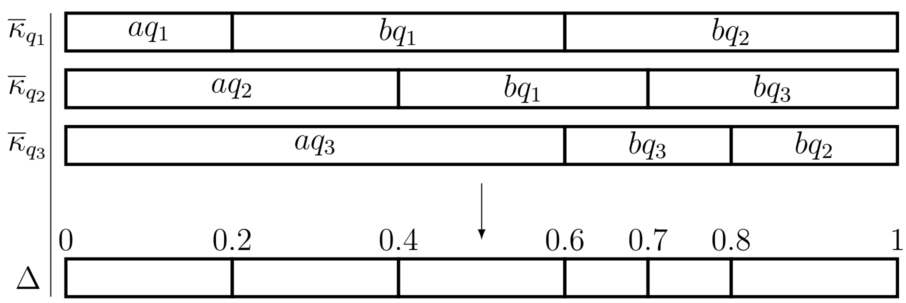
$$\begin{aligned}
\mathbb{P}_{e_{r_0}}(a_1 r_1 w (\Sigma Q)^\omega) &= M(a_1)_{r_0,r_1} \mathbb{P}_{e_{r_1}}(w(\Sigma Q)^\omega) \\
&= \mathbb{P}_\rho\Big(\kappa_{r_0}^{-1}(\{a_1 r_1\})\Delta^\omega\Big)\mathbb{P}_\rho\Big(\kappa_{r_1}^{-1}(\{w\})\Delta^\omega\Big) \\
&= \mathbb{P}_\rho\Big(\kappa_{r_0}^{-1}(\{a_1 r_1\})\kappa_{r_1}^{-1}(\{w\})\Delta^\omega\Big) \\
&= \mathbb{P}_\rho\Big(\kappa_{r_0}^{-1}(\{a_1 r_1 w\})\Delta^\omega\Big)
\end{aligned}$$

by the independence of $\mathbb{P}_\rho$. This completes the induction. $\square$

Example 7.2.14. We demonstrate the construction of Δ from the proof of Lemma 7.2.13 using the example HMM below.



In the diagram below we give a representation of the functions $\overline{\kappa}_{q_1}, \overline{\kappa}_{q_2}, \overline{\kappa}_{q_3}$ and also the resulting set Δ. The first three horizontal stacks of rectangles each represent a partition of the interval $[0,1)$ into the values taken by $\overline{\kappa}_{q_i}$ for $i = 1,2,3$. The bottom horizontal stack of rectangles represent the intervals in Δ.



We have $\Delta = \{[0,0.2),[0.2,0.4),[0.4,0.6),[0.6,0.7),[0.7,0.8),[0.8,1)\}$ and κ_{q_1} is defined piecewise by

$$\kappa_{q_1}(x) = \begin{cases} aq_1 & x = [0,0.2) \\ bq_1 & x \in \{[0.2,0.4),[0.4,0.6)\} \\ bq_2 & x \in \{[0.6,0.7),[0.7,0.8),[0.8,1)\}. \end{cases} \tag{7.5}$$

Proof of Lemma 7.2.12. By Lemma 7.2.13, since (Q_2,Σ,M_2) is a HMM, we may compute in polynomial time a finite set Δ, a distribution $\rho \in [0,1]^\Delta$ and for each $r \in Q_2$ a mapping $\kappa_r : \Delta \to \Sigma Q_2$ with the property stated in Lemma 7.2.13. Then, we define $\tilde{M} : \Delta \to [0,1]^{Q_1\times Q_2}$ such that for all $(q_0,r_0),(q,r) \in Q_1 \times Q_2$:

$$\tilde{M}(\overline{a})_{(q_0,r_0),(q,r)} = \begin{cases} M_1(a)_{q_0,q} & \text{when } \kappa_{r_0}(\overline{a}) = ar \\ 0 & \text{otherwise.} \end{cases}$$

We extend $\tilde{M}$ to the mapping $\tilde{M} : \Delta^* \to [0,1]^{(Q_1\times Q_2)\times(Q_1\times Q_2)}$ by $\tilde{M}(\overline{a}_1\cdots\overline{a}_n) = \tilde{M}(\overline{a}_1)\cdots\tilde{M}(\overline{a}_n)$.

Let $r_0 \in Q_2$, $\overline{a}_1\cdots\overline{a}_n \in \Delta^*$ and $\kappa_{r_0}(\overline{a}_1\cdots\overline{a}_n) = a_1r_1\cdots a_nr_n$. By the definition of the extension of κ_{r_0}, it follows that $\wedge_{i=1}^n \kappa_{r_{i-1}}(\overline{a}_i) = a_ir_i$. For all $q_0,\ldots q_n \in Q_1$ we have

$$\begin{aligned} M_1(a_1)_{q_0,q_1}\cdots M_1(a_n)_{q_{n-1},q_n} &= \tilde{M}(\overline{a}_1)_{(q_0,r_0),(q_1,r_1)}\cdots\tilde{M}(\overline{a}_n)_{(q_{n-1},r_{n-1}),(q_n,r_n)} \\ &= \sum_{\tilde{r}_1,\ldots,\tilde{r}_n\in Q_2} \tilde{M}(\overline{a}_1)_{(q_0,\tilde{r}_0),(q_1,\tilde{r}_1)}\cdots\tilde{M}(\overline{a}_n)_{(q_{n-1},\tilde{r}_{n-1}),(q_n,\tilde{r}_n)} \end{aligned}$$

and therefore by the definition of $\tilde{M}$,

$$\begin{aligned} &\|(\pi_1\times e_{r_0})\tilde{M}(\overline{a}_1\cdots\overline{a}_n)\| \\ &= \sum_{q_0\in\mathrm{supp}\,(\pi_1)}\ \sum_{(q_1,\tilde{r}_1),\ldots(q_n,\tilde{r}_n)\in C} (\pi_1)_{q_0}\tilde{M}(\overline{a}_1)_{(q_0,\tilde{r}_0),(q_1,\tilde{r}_1)}\cdots\tilde{M}(\overline{a}_n)_{(q_{n-1},\tilde{r}_{n-1}),(q_n,\tilde{r}_n)} \\ &= \sum_{q_0\in\mathrm{supp}\,(\pi_1)}\ \sum_{q_1,\ldots q_n\in Q_1}\ \sum_{\tilde{r}_1,\ldots\tilde{r}_n\in Q_2} (\pi_1)_{q_0}\tilde{M}(\overline{a}_1)_{(q_0,\tilde{r}_0),(q_1,\tilde{r}_1)}\cdots\tilde{M}(\overline{a}_n)_{(q_{n-1},\tilde{r}_{n-1}),(q_n,\tilde{r}_n)} \\ &= \sum_{q_0\in\mathrm{supp}\,(\pi_1)}\ \sum_{q_1,\ldots q_n\in Q_1} (\pi_1)_{q_0}\tilde{M}(\overline{a}_1)_{(q_0,r_0),(q_1,r_1)}\cdots\tilde{M}(\overline{a}_n)_{(q_{n-1},r_{n-1}),(q_n,r_n)} \\ &= \sum_{q_0\in\mathrm{supp}\,(\pi_1)}\ \sum_{q_1,\ldots q_n\in Q_1} (\pi_1)_{q_0}M_1(a_1)_{q_0,q_1}\cdots M_1(a_n)_{q_{n-1},q_n} \\ &= \|\pi_1M_1(a_1\cdots a_n)\|. \end{aligned} \tag{7.6}$$

Suppose there is an edge in the joint graph $G_{(Q_1,\Sigma,M_1),(Q_2,\Sigma,M_2)}$ between two states $(q_0,r_0),(q,r) \in C$ then there is $a \in \Sigma$ such that $M_1(a)_{q_0,q} > 0$ and $M_2(a)_{r_0,r} > 0$.

Hence, $\mathbb{P}_{e_{r_0}}(a(\Sigma Q)^\omega) > 0$ and so by Lemma 7.2.13 there exists $\overline{a} \in \kappa_{r_0}^{-1}(\{ar\})$. It follows that $\tilde{M}(\overline{a})_{(q_0,r_0),(q,r)} > 0$. Therefore, since C is a bottom SCC of the joint graph $G_{(Q_1,\Sigma,M_1),(Q_2,\Sigma,M_2)}$, we have that the graph of the matrix system $(C, \Delta, \tilde{M}_{|C})$ is strongly connected and therefore $((C, \Delta, \tilde{M}_{|C}), \rho)$ is a Lyapunov system. Moreover, for any $\pi_1 \in [0,1]^{Q_1}$ and $\pi_2 \in [0,1]^{Q_2}$ such that $\mathrm{supp}(\pi_1) \times \mathrm{supp}(\pi_2) \subseteq C$ we have that $\|(\pi_1 \times \pi_2)\tilde{M}(\overline{a}_1 \cdots \overline{a}_n)\| = \|(\pi_1 \times \pi_2)_{|C}\tilde{M}_{|C}(\overline{a}_1 \cdots \overline{a}_n)\|$ where $(\pi_1 \times \pi_2)_{|C}$ is the truncation of the vector $\pi_1 \times \pi_2$ to C. Further, we have for all $x \in \mathbb{R}$ and $n \in \mathbb{N}$

$$
\begin{aligned}
&\mathbb{P}_{\pi_2}(\{w \in \Sigma^\omega \mid \|\pi_1 M(w_n)\| = x\}).\\
&= \sum_{r_0 \in \mathrm{supp}\,(\pi_2)} (\pi_2)_{r_0} \mathbb{P}_{e_{r_0}}(\{w \in \Sigma^\omega \mid \|\pi_1 M_1(w_n)\| = x\})\\
&= \sum_{r_0 \in \mathrm{supp}\,(\pi_2)} (\pi_2)_{r_0} \sum_{a_1 r_1 \cdots a_n r_n \in (\Sigma Q_2)^n \text{ s.t. } \|\pi_1 M_1(a_1 \cdots a_n)\| = x} \mathbb{P}_{e_{r_0}}(a_1 r_1 \cdots a_n r_n (\Sigma Q)^\omega)\\
&= \sum_{r_0 \in \mathrm{supp}\,(\pi_2)} (\pi_2)_{r_0} \sum_{a_1 r_1 \cdots a_n r_n \in (\Sigma Q_2)^n \text{ s.t. } \|\pi_1 M_1(a_1 \cdots a_n)\| = x} \mathbb{P}_{\rho}(\kappa_{r_0}^{-1}(a_1 r_1 \cdots a_n r_n)(\Delta)^\omega)\\
&\quad \text{by Lemma 7.2.13}\\
&= \sum_{r_0 \in \mathrm{supp}\,(\pi_2)} (\pi_2)_{r_0} \mathbb{P}_{\rho}(\{\overline{w} \in \Delta^\omega \mid \kappa_{r_0}(\overline{w}_n) = a_1 r_1 \cdots a_n r_n,\ \|\pi_1 M_1(a_1 \cdots a_n)\| = x\})\\
&= \sum_{r_0 \in \mathrm{supp}\,(\pi_2)} (\pi_2)_{r_0} \mathbb{P}_{\rho}(\{\overline{w} \in \Delta^\omega \mid \|(\pi_1 \times e_{r_0})\tilde{M}(\overline{w}_n)\| = x\}) \quad \text{by (7.6)}\\
&= \sum_{r_0 \in \mathrm{supp}\,(\pi_2)} (\pi_2)_{r_0} \mathbb{P}_{\rho}(\{\overline{w} \in \Delta^\omega \mid \|(\pi_1 \times e_{r_0})_{|C}\tilde{M}_{|C}(\overline{w}_n)\| = x\}).
\end{aligned}
\tag{7.7}
$$

Combining the above equalities with Lemma 7.2.2 we have that there exists a $\lambda \in [-\infty, 0]$ which does not depend on π_1 or π_2 such that

$$
\begin{aligned}
1 &= \sum_{r_0 \in \mathrm{supp}\,\pi_2} (\pi_2)_{r_0}\\
&= \sum_{r_0 \in \mathrm{supp}\,\pi_2} (\pi_2)_{r_0} \mathbb{P}_{\rho}\Big(\Big\{\overline{w} \in \Delta^\omega \mid\\
&\qquad \lim_{n\to\infty} \frac{1}{n} \ln \|(\pi_1 \times e_{r_0})_{|C}\tilde{M}_{|C}(\overline{w}_n)\| \in \{-\infty, \lambda\}\Big\}\Big) \text{ by Lemma 7.2.2}\\
&= \mathbb{P}_{\pi_2}\Big(\Big\{w \in \Sigma^\omega \mid \lim_{n\to\infty} \frac{1}{n} \ln \|\pi_1 M(w_n)\| \in \{-\infty, \lambda\}\Big\}\Big) \text{ by (7.7).}
\end{aligned}
$$

□

7.2.4 Right-Bottom SCCs

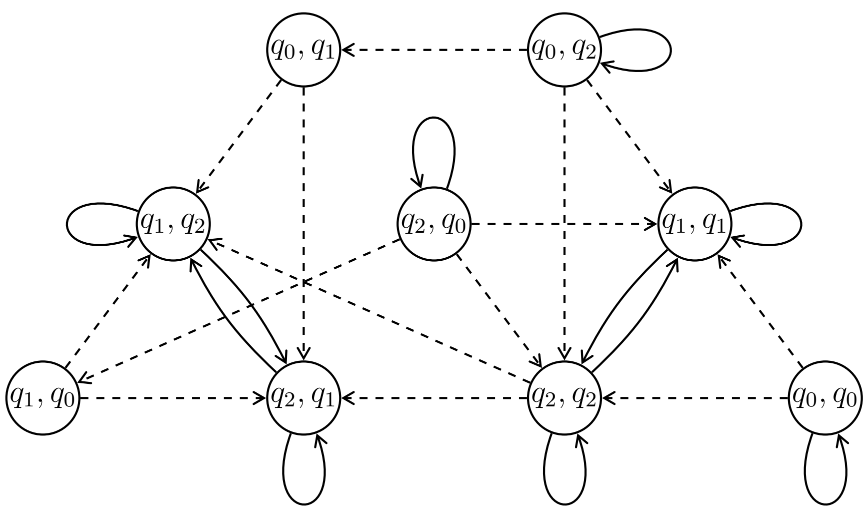
Let $\mathcal{F} = (Q, \Sigma, M)$ be a HMM. Let $R \subseteq Q \times Q$ be a (not necessarily bottom) SCC of the joint graph $G_{\mathcal{F},\mathcal{F}}$ such that $Q_R := \{q_2 \in Q \mid \exists\, q_1 \in Q : (q_1, q_2) \in R\}$ is a bottom SCC of the embedded graph of $\mathcal{F}$. We call such R a *right-bottom* SCC. Clearly there are at most $|Q|^2$ right-bottom SCCs. Towards Theorem 7.2.4 we want to define, for each right-bottom SCC R, two generalized Lyapunov systems $\mathcal{S}^1_R, \mathcal{S}^2_R$. Intuitively, $\mathcal{S}^1_R$ and $\mathcal{S}^2_R$ correspond to the numerator and the denominator of the likelihood ratio, respectively.

Recall the observation density matrix M' defined in Section 7.2.3. Since Q_R is a bottom SCC of the embedded graph of $\mathcal{F}$, the HMM $\mathcal{F}_R := (Q_R, \Sigma \times Q_R, M'_{|Q_R})$ is strongly connected. This HMM $\mathcal{F}_R$ will be used both in $\mathcal{S}^1_R$ and in $\mathcal{S}^2_R$. Next, define $\overline{M} : (\Sigma \times Q) \to [0,1]^{(Q\times Q)\times(Q\times Q)}$ by

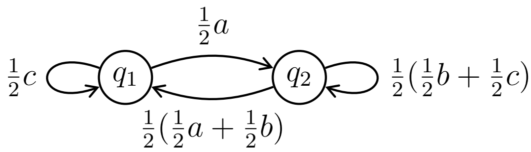
$$\overline{M}(a, r_2)_{(q_1,q_2),(r_1,r_2)} \; := \; M(a)_{q_1,r_1} \quad \text{for all } a \in \Sigma \; \text{ and } q_1, q_2, r_1, r_2 \in Q\,.$$

Now define $\mathcal{S}^1_R := (\mathcal{M}^1, \mathcal{F}_R, C^1)$, where $\mathcal{M}^1 := (R, \Sigma \times Q_R, \overline{M}_{|R})$ and $C^1 := \{((q_1,q_2),q_2) \mid (q_1,q_2) \in R\}$. Finally, denoting by $R' \subseteq Q_R \times Q_R$ the SCC of the joint graph $G_{\mathcal{F},\mathcal{F}}$ that contains the "diagonal" vertices $(q,q) \in Q_R \times Q_R$, define $\mathcal{S}^2_R := (\mathcal{M}^2, \mathcal{F}_R, C^2)$, where $\mathcal{M}^2 := (R', \Sigma \times Q_R, \overline{M}_{|R'})$ and $C^2 := \{((q_1,q_2),q_2) \mid (q_1,q_2) \in R'\}$.

Example 7.2.15. Recall the matrix system $\mathcal{M} = (Q_1, \Sigma, M_1)$ from Example 7.2.11. Since M_1 is stochastic, $\mathcal{M}$ is also a HMM. Let $\mathcal{F} = \mathcal{M}$ then the graph $G_{\mathcal{F},\mathcal{F}}$ is shown below. To help dissect the diagram, observe that we have taken the graph from Example 7.2.11 and appended the states (q_0,q_0), (q_1,q_0) and (q_2,q_0). We once again use dashed lines to represent edges between SCCs.

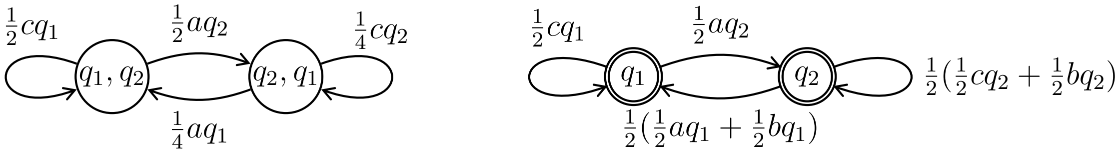


The right-bottom SCCs of $G_{\mathcal{F},\mathcal{F}}$ are $\{(q_1,q_1),(q_2,q_2)\}$ and $\{(q_1,q_2),(q_2,q_1)\}$. For any of these right-bottom SCCs, $Q_R = \{q_1, q_2\}$ and $\mathcal{F}_R$ is shown below.

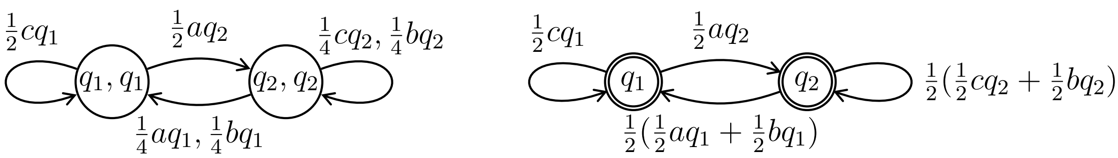


Note that $\mathcal{F}_R$ is equivalent to the HMM from Example 4.1.1.

For $R = \{(q_1,q_2),(q_2,q_1)\}$ we have $C^1 = \{((q_1,q_2),q_2),(q_2,q_1),q_1)\}$ and the generalized Lyapunov system $\mathcal{S}^1_R$ is shown below.



We have $R' = \{(q_1,q_1),(q_2,q_2)\}$ and $C^2 = \{((q_1,q_1),q_1),(q_2,q_2),q_2)\}$ and so the generalized Lyapunov system $\mathcal{S}^2_R$ is shown below.



For $R = \{(q_1,q_1),(q_2,q_2)\}$ we have $R' = R$ and so $\mathcal{S}^1_R = \mathcal{S}^2_R$.

We are ready to state the following key technical lemma:

Lemma 7.2.16. *Given a HMM (Q,Σ,M), let $\mathcal{R} \subseteq 2^{Q\times Q}$ be the set of its right-bottom SCCs, and, for $R \in \mathcal{R}$, let $\mathcal{S}^1_R, \mathcal{S}^2_R$ be the generalized Lyapunov systems defined above. Then, for any initial distributions π_1, π_2, the limit $\lim_{n\to\infty} \frac{1}{n} \ln L_n$ exists $\mathbb{P}_{\pi_2}$-a.s. and lies in*

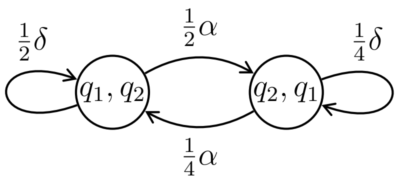
$$\{-\infty\} \cup \{\lambda(\mathcal{S}^1_R) - \lambda(\mathcal{S}^2_R) \mid R \in \mathcal{R},\ \mathit{supp}(\pi_1) \times \mathit{supp}(\pi_2) \longrightarrow_{G_{\mathcal{F},\mathcal{F}}} R\}\,.$$

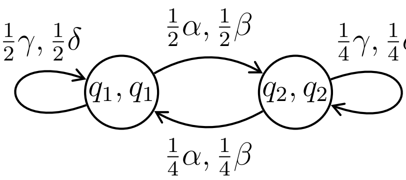
Thus, $\Lambda_{\pi_1,\pi_2} \subseteq \{-\infty\} \cup \{\lambda(\mathcal{S}^1_R) - \lambda(\mathcal{S}^2_R) \mid R \in \mathcal{R},\ \mathit{supp}(\pi_1) \times \mathit{supp}(\pi_2) \longrightarrow_{G_{\mathcal{F},\mathcal{F}}} R\}$.

Example 7.2.17. Consider the HMM $\mathcal{F}$ from Example 7.2.15 with initial distributions $\pi_1 = e_{q_0}$ and $\pi_2 = \{e_{q_2}\}$. Both right-bottom SCCs $\{(q_1, q_1), (q_2, q_2)\}$ and $\{(q_1, q_2), (q_2, q_1)\}$ are reachable from (q_0, q_2) in $G_{\mathcal{F},\mathcal{F}}$. We have $\mathcal{S}^1_{\{(q_1,q_1),(q_2,q_2)\}} = \mathcal{S}^2_{\{(q_1,q_1),(q_2,q_2)\}}$ and hence

$$\lim_{n\to\infty} \frac{1}{n} \ln L_n \in \left\{-\infty, \lambda\left(\mathcal{S}^1_{\{(q_1,q_2),(q_2,q_1)\}}\right) - \lambda\left(\mathcal{S}^2_{\{(q_1,q_2),(q_2,q_1)\}}\right), 0\right\}.$$

Using Lemma 7.2.12 we compute $\Delta = \{\alpha, \beta, \gamma, \delta\}$, $\rho = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and for $\mathcal{S}^1_{\{(q_1,q_2),(q_2,q_1)\}}$ and $\mathcal{S}^2_{\{(q_1,q_2),(q_2,q_1)\}}$ the matrix systems M_1 and M_2 respectively, shown left and right.





These matrix systems are the restrictions of M from Example 7.2.5 to $\{(q_1, q_1), (q_2, q_2)\}$ and $\{(q_1, q_2), (q_2, q_1)\}$ respectively. Clearly $\lambda((\{(q_1, q_2), (q_2, q_1)\}, \Delta, M_1), \rho) = -\infty$ since any word with a β or γ is mortal for M_1. Therefore, $\lim_{n\to\infty} \frac{1}{n} \ln L_n \in \{-\infty, 0\}$.

Proof of Lemma 7.2.16. Let $\mathcal{F} = (Q, \Sigma, M)$ and observe that

$$\mathbb{P}_{\pi_2}\Big(r_0 a_1 r_1 a_2 r_2 \cdots \in Q(\Sigma Q)^\omega \mid \exists k \in \mathbb{N} \text{ s.t. } r_k \text{ is in a bottom SCC of } \mathcal{F}\Big) = 1.$$

Consider a word $a_1 r_1 a_2 r_2 \cdots \in (\Sigma Q)^\omega$ such that $\mathbb{P}_{\pi_2}(Q a_1 r_1 \cdots a_m r_m (\Sigma Q)^\omega) > 0$ for all $m \in \mathbb{N}$ and further, there exists a $k \in \mathbb{N}$ such that $r_k \in P$ where $P \subseteq Q$ is a bottom SCC of $\mathcal{F}$. We write $u = a_1 r_1 \cdots a_k r_k$ and $w = a_{k+1} r_{k+1} a_{k+2} r_{k+2} \cdots \in (\Sigma Q)^\omega$. Let $\mu_i = \pi_i M(u) \times e_{r_k}$ for $i = 1, 2$. Recall the definition of $\overline{M}$. We have that for any $n > k$ that $\|\pi_i M(a_1 \cdots a_{k+n})\| = \|\pi_i M(a_1 \cdots a_k) M(a_{k+1} \cdots a_{k+n})\| = \|\mu_i \overline{M}(w_n)\|$ for $i = 1, 2$. We fix u and consider words w produced by $\mathcal{F}$ with initial distribution e_{r_k}. We have that $\mathbb{P}_{e_{r_k}}$-almost surely if the limits exist,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{k+n} \ln L_{k+n}(uw) &= \lim_{n\to\infty} \frac{1}{k+n} \Big[\ln \|\pi_1 M(a_1 \cdots a_{k+n})\| - \ln \|\pi_2 M(a_1 \cdots a_{k+n})\|\Big] \\
&= \lim_{n\to\infty} \frac{1}{k+n} \Big[\ln \|\mu_1 \overline{M}(w_n)\| - \ln \|\mu_2 \overline{M}(w_n)\|\Big] \\
&= \lim_{n\to\infty} \frac{1}{n} \Big[\ln \|\mu_1 \overline{M}(w_n)\| - \ln \|\mu_2 \overline{M}(w_n)\|\Big]
\end{aligned}$$

since $\lim_{n\to\infty} \frac{n}{k+n} = 1$.

By Lemma 7.2.13 there is a finite set Δ, distribution $\rho \in (0,1]^\Delta$ and mapping $\kappa_{r_k} : \Delta^+ \to (\Sigma Q)^+$ such that $\mathbb{P}_\rho(\{\kappa_{r_k}^{-1}(w_n)\}\Delta^\omega) = \mathbb{P}_{e_{r_k}}(w_n(\Sigma Q)^\omega)$ for all $n \in \mathbb{N}$. Let $S \subseteq Q \times Q$. It follows that for each $n \in \mathbb{N}$ and $A \in [0,1]^{S\times S}$ we have

$$\mathbb{P}_{e_{r_k}}(\{w \in (\Sigma Q)^\omega \mid \overline{M}_{|S}(w_n) = A\}) = \mathbb{P}_\rho(\{\overline{w} \in \Delta^\omega \mid \overline{M}_{|S}(\kappa_{r_k}(\overline{w}_n)) = A\}). \quad (7.8)$$

Let $\mathcal{C}$ be the set of SCCs of $G_{\mathcal{F},\mathcal{F}}$. We define $\sigma : \mathcal{C} \to \mathcal{P}(Q)$ such that $\sigma(R) = \{q \in Q \mid \exists p \in Q \text{ s.t. } (p,q) \in R\}$. Recall that $r_k \in P$ and P is a bottom SCC. Thus, for all $i = 1, 2$, $n \in \mathbb{N}$ and $(p,q) \in \operatorname{supp}(\mu_i \overline{M}(w_n))$ we have $q \in P$. Therefore, for all $R \in \mathcal{C}$

$$\operatorname{supp}(\mu_i) \to_{G_{\mathcal{F},\mathcal{F}}} R \implies \sigma(R) = P. \quad (7.9)$$

Consider the composition $\overline{M} \circ \kappa_{r_k} : \Delta^+ \to [0,1]^{(Q\times Q)\times(Q\times Q)}$. For any $R \in \sigma^{-1}(P)$ we have that trivially

$$(\overline{M} \circ \kappa_{r_k})_{|R} = \overline{M}_{|R} \circ \kappa_{r_k}. \quad (7.10)$$

Further, $((R, \Delta, \overline{M}_{|R} \circ \kappa_{r_k}), \rho)$ is a Lyapunov system and recall that

$$S_R^1 = ((R, \Sigma P, \overline{M}_{|R}), (P, \Sigma P, M'_{|P}), C_R^1)$$

is a generalized Lyapunov system where $C_R^1 = \{((q_1,q_2),q_2) \mid (q_1,q_2) \in R\}$. Let $V = \{q \in Q \mid (q, r_k) \in R\}$ and $\upsilon \in [0,1]^V$. Since $\operatorname{supp}(\upsilon \times e_{r_k}) \times \{r_k\} \in C_R^1$ by Lemma 7.2.12 the limit $\lim_{n\to\infty} \frac{1}{n} \ln \|(\upsilon \times e_{r_k})\overline{M}_{|R}(w_n)\|$ exists $\mathbb{P}_{e_{r_k}}$-almost surely and equals either $\lambda(S_R^1)$ or $-\infty$. Then, by (7.8) with $S = R$ we have for all $x \in [-\infty, 0]$ that

$$\begin{aligned}&\mathbb{P}_{e_{r_k}}(\{w \in (\Sigma Q)^\omega \mid \lim_{n\to\infty} \frac{1}{n} \ln \|(\upsilon \times e_{r_k})\overline{M}_{|R}(w_n)\| = x\}) \\ &= \mathbb{P}_\rho(\{\overline{w} \in \Delta^\omega \mid \lim_{n\to\infty} \frac{1}{n} \ln \|(\upsilon \times e_{r_k})\big[(\overline{M}_{|R} \circ \kappa_{r_k})(\overline{w}_n)\big]\| = x\}). \quad (7.11)\end{aligned}$$

Recall the definition of $\mathcal{L}$. The following series of equalities hold:

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \ln \|\mu_i \overline{M}(w_n)\| \quad \mathbb{P}_{e_{r_k}}\text{-a.s.} \\
&= \lim_{n\to\infty} \frac{1}{n} \ln \|\mu_i \Big[(\overline{M} \circ \kappa_{r_k})(\overline{w}_n)\Big]\| \quad \mathbb{P}_\rho\text{-a.s.} \quad \text{by (7.8) where } S = Q \times Q \\
&\in \Big\{\lambda((R, \Delta, (\overline{M} \circ \kappa_{r_k})_{|R}), \rho) \mid R \in \mathcal{C}/\mathcal{L}, \operatorname{supp}(\mu_i) \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\} \cup \{-\infty\}\, \mathbb{P}_\rho\text{-a.s.} \\
&\qquad \text{by Lemma 7.2.9} \\
&= \Big\{\lambda((R, \Delta, \overline{M}_{|R} \circ \kappa_{r_k}), \rho) \mid R \in \sigma^{-1}(P)/\mathcal{L}, \operatorname{supp}(\mu_i) \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\} \cup \{-\infty\} \\
&\qquad \text{by (7.9) and (7.10)} \\
&= \Big\{\lambda(\mathcal{S}^1_R) \mid R \in \sigma^{-1}(P)/\mathcal{L}, \operatorname{supp}(\mu_i) \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\} \cup \{-\infty\} \quad \text{by (7.11).} \\
&= \Big\{\lambda(\mathcal{S}^1_R) \mid R \in \sigma^{-1}(P)/\mathcal{L}, \operatorname{supp}(\mu_i) \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\} \cup \{-\infty\}.
\end{aligned} \tag{7.12}$$

We now focus on the case $i = 2$. Since P is an SCC, there is a unique right-bottom SCC P' that contains the states $\{(p, p) \mid p \in P\}$. Let U' be another right-bottom SCC in the set $\sigma^{-1}(P)/\mathcal{L}$. Since U' is an SCC and not in $\mathcal{L}$ we have that $\| \sum_{(s,r)\in U'} (e_s \times e_r)\overline{M}_{|U'}(w_n)\| > 0$ for all $n \in \mathbb{N}$. Therefore $\mathbb{P}_{e_r}$-a.s. by Lemma 7.2.12,

$$\begin{aligned}
\lambda(\mathcal{S}^1_{U'}) &= \lim_{n\to\infty} \frac{1}{n} \ln \| \sum_{(s,r)\in U'} (e_s \times e_r)\overline{M}_{|U'}(w_n)\| \\
&\le \lim_{n\to\infty} \frac{1}{n} \ln |U'| \max_{(s,r)\in U'} \|(e_s \times e_r)\overline{M}_{|U'}(w_n)\| \\
&= \max_{(s,r)\in U'} \lim_{n\to\infty} \frac{1}{n} \ln \|(e_s \times e_r)\overline{M}_{|U'}(w_n)\| \\
&= \lambda(\mathcal{S}^1_{U'}) \text{ because each limit is either } \lambda(\mathcal{S}^1_{U'}) \text{ or } -\infty.
\end{aligned}$$

It follows that there is some $(s, r) \in U'$ such that $\mathbb{P}_{e_r}(\lim_{n\to\infty} \frac{1}{n} \ln \|(e_s \times e_r)\overline{M}_{|U'}(w_n)\| = \lambda(\mathcal{S}^1_{U'})) > 0$. Let $U = \{q \in Q \mid \exists p \in Q \text{ s.t. } (q, p) \in U'\}$. Then $s \in U$, $r \in P$ and $(r, r) \in P'$. Hence, with $\mathbb{P}_{e_r}$-probability greater than 0 we have

$$\begin{aligned}
0 &\ge \lim_{n\to\infty} \frac{1}{n} \ln \frac{\|e_s M_{|U}(w_n)\|}{\|e_r M_{|P}(w_n)\|} \text{ by Lemma 6.0.1} \\
&= \lim_{n\to\infty} \frac{1}{n} \ln \frac{\|(e_s \times e_r)\overline{M}_{|U'}(w_n)\|}{\|(e_r \times e_r)\overline{M}_{|P'}(w_n)\|} \\
&= \lim_{n\to\infty} \frac{1}{n} \Big[\ln \|(e_s \times e_r)\overline{M}_{|U'}(w_n)\| - \ln \|(e_r \times e_r)\overline{M}_{|P'}(w_n)\|\Big] \\
&= \lambda(\mathcal{S}^1_{U'}) - \lambda(\mathcal{S}^1_{P'})
\end{aligned}$$

which implies that $\lambda(\mathcal{S}^1_{P'}) \geq \lambda(\mathcal{S}^1_{U'})$. We have that $(r_k, r_k) \in \mathrm{supp}\,(\mu_2)$ and so $\mathbb{P}_{e_{r_k}}$-almost surely we have $\|(e_{r_k} \times e_{r_k})\overline{M}_{P'}(w_n)\| > 0$ for all $n \in \mathbb{N}$. By (7.9), it follows that $\mathbb{P}_{e_{r_k}}$-a.s. we have

$$\begin{aligned}
\lambda(\mathcal{S}^1_{P'}) &= \lim_{n\to\infty} \frac{1}{n} \ln \|(e_{r_k} \times e_{r_k})\overline{M}_{|P'}(w_n)\| \quad \text{by Lemma 7.2.12} \\
&\leq \lim_{n\to\infty} \frac{1}{n} \ln \|\mu_2 \overline{M}(w_n)\| \\
&\leq \max\Big\{\lambda(\mathcal{S}^1_R) \mid R \in \sigma^{-1}(P),\ \mathrm{supp}\,(\mu_2) \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\} \text{ by (7.12)} \\
&\leq \lambda(\mathcal{S}^1_{P'}).
\end{aligned}$$

Then, recalling the definition of S^2_R we have $\mathbb{P}_{\pi_2}$-almost surely,

$$\begin{aligned}
\lim_{n\to\infty} \frac{\ln L_{k+n}(uw)}{k+n} &= \lim_{n\to\infty} \frac{1}{n}\Big[\ln \|\mu_1 \overline{M}(w_n)\| - \ln \|\mu_2 \overline{M}(w_n)\|\Big] \\
&\in \Big\{\lambda(\mathcal{S}^1_R) - \lambda(S^1_{P'}) \mid R \in \sigma^{-1}(P), \mathrm{supp}\,(\mu_1) \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\} \cup \{-\infty\} \\
&= \Big\{\lambda(\mathcal{S}^1_R) - \lambda(S^2_R) \mid R \in \sigma^{-1}(P), \mathrm{supp}\,(\mu_1) \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\} \cup \{-\infty\}
\end{aligned}$$

since $\mathcal{S}^2_R = \mathcal{S}^1_{P'}$ when $\sigma(R) = P$.

We may now relax the dependence on u and P. We have that $\mathrm{supp}\,(\pi_1) \times \mathrm{supp}\,(\pi_2) \to_{G_{\mathcal{F},\mathcal{F}}} (\mathrm{supp}\,(\pi_1 M(a_1 \cdots a_k))) \times \{r_k\}$ for any $a_1 r_1 \cdots a_k r_k \in (\Sigma Q)^+$ where $\mathbb{P}_{\pi_2}(Q a_1 r_1 \cdots a_k r_k (\Sigma Q)^\omega) > 0$. Therefore for all $R \in \mathcal{R}$

$$(\mathrm{supp}\,(\pi_1 M(a_1 \cdots a_k))) \times \{r_k\} \to_{G_{\mathcal{F},\mathcal{F}}} R \implies \mathrm{supp}\,(\pi_1) \times \mathrm{supp}\,(\pi_2) \to_{G_{\mathcal{F},\mathcal{F}}} R.$$

Finally we have up to a $\mathbb{P}_{\pi_2}$-null set

$$\begin{aligned}
&\{a_1 r_1 a_2 r_2 \cdots \in (\Sigma Q)^\omega \mid \exists k \in \mathbb{N} \text{ s.t. } r_k \text{ is in a bottom SCC of } \mathcal{F}\} \\
&\subseteq \Big\{ \lim_{n\to\infty} \frac{1}{k+n} \ln L_{k+n}(uw) \\
&\in \{-\infty\} \cup \Big\{\lambda(\mathcal{S}^1_R) - \lambda(\mathcal{S}^2_R) \mid R \in \mathcal{R}, \mathrm{supp}\,(\pi_1) \times \mathrm{supp}\,(\pi_2 \to_{G_{\mathcal{F},\mathcal{F}}} R\Big\}\Big\}
\end{aligned}$$

and the lemma follows. □

With Lemma 7.2.16 at hand, the proof of Theorem 7.2.4 is easy:

Proof of Theorem 7.2.4. As argued before, the set $\mathcal{R}$ of right-bottom SCCs of the given HMM has at most $|Q|^2$ elements. These right-bottom SCCs R and the associated generalized Lyapunov systems $\mathcal{S}^1_R, \mathcal{S}^2_R$ can be computed in polynomial time. By Lemma 7.2.16 we have $\Lambda = \bigcup_{\pi_1,\pi_2} \Lambda_{\pi_1,\pi_2} \subseteq \{-\infty\} \cup \{\lambda(\mathcal{S}^1_R) - \lambda(\mathcal{S}^2_R) \mid R \in \mathcal{R}\}$. By the second point of Lemma 7.2.12, for each $R \in \mathcal{R}$ one can compute in polynomial time an equivalent Lyapunov system. □

8 Conclusions

8.1 Summary

We have shown that equivalence of continuous-observation HMMs is decidable in polynomial time, by reduction to the finite-observation case. The crucial insight is that, rather than integrating the density functions, one needs to consider them as elements of a vector space and computationally establish linear (in)dependence of functions. Therefore, our polynomial-time reduction performs symbolic computations on continuous density functions. As a suitable framework for these computations we have introduced the notion of linearly decomposable profile languages, and we have established Γ_{GEM} as such a profile language.

We then built on this topic by showing that distinguishability of continuous-observation HMMs is decidable in polynomial time under a slightly modified framework for the computation. We require that a HMM is encoded with a linearly decomposable and joint-graph decomposable profile language. A HMM is encoded with a joint-graph decomposable profile language if, for every pair of transitions, we can decide whether both transitions can happen simulataneously with non-zero probability. This condition is considerably easier to deal with than linear decomposability: most standard probability distribution functions have predefined supports which don't depend on their parameters. Once again, our polynomial-time algorithm is a symbolic computation using density functions.

Next, we demonstrated general techniques for establishing whether a profile language has the right properties for the equivalence and distinguishability algorithms. Specifically, we defined a "large" profile language Span-$\mathcal{I}(\Gamma_m \cup \Gamma_t \cup \Gamma_\chi \cup \Gamma_e \cup \Gamma_{mb} \cup \Gamma_g)$ and showed that it is linearly decomposable and joint-graph decomposable. We also defined a profile language to encode Dirichlet distribution pdfs and showed that this profile language is linearly decomposable and joint-graph decomposable. This example is interesting because Dirichlet random variables are multi-dimensional.

We have shown that two initial distributions may be distinguished using the sequential probability ratio test. The execution time of the SPRT is tightly connected with likelihood exponents. These numbers are related to Lyapunov exponents and can be viewed as a distance measure between HMMs. We have shown that the number of likelihood exponents is quadratic in the number of states. The associated computational problems tend to be complex (PSPACE-hard), but become tractable for deterministic HMMs. In our work we did not make any ergodicity assumptions on the HMMs, unlike in earlier works from mathematics and engineering such as [48–51].

8.2 Open Questions

It remains to show results for the SPRT and Likelihood exponents in the case of continuous-observation HMMs. We could try and show the following hold for continuous-observation HMMs:

1. Lemma 6.0.1 and Proposition 6.1.1 and so the SPRT terminates almost-surely,

2. The Wald formula Theorem 6.2.4 and its almost-sure variants Proposition 6.3.6, Proposition 6.3.2 and Proposition 6.3.5,

3. Theorem 7.2.4, thus there are at most $|Q|^2 + 1$ likelihood exponents.

In order to prove these, we would need continuous-observation variants of some results from the literature.

1. Proposition 6 from [23] shows that the likelihood ratio converges almost-surely when the two initial distributions are distinguishable.

2. Theorem 5 and Lemma 9 from [24] provide the backbone for the Wald formula generalisation.

3. Proposition 1 from [45] shows that a Lyapunov system defines a single Lyapunov exponent. This result is crucial for the upper bound on likelihood exponents.

In addition, we might try to generalise the computational parts of Chapter 6 and Chapter 7. Towards an analogue of Theorem 6.4.1 we could define a Numerator Support Markov Chain for continuous-observation HMMs. For $S \subseteq Q$ and $x \in \Sigma$, define $\delta(S,x) := \{q' \in Q \mid \exists\, q \in S : \Psi(x)_{q,q'} > 0\}$. Define the Markov chain $\mathcal{D} := (2^Q \times Q, T)$ where

$$T_{(S,q),(S',q')} := \int_{\delta(S,x)=S'} \Psi(x)_{q,q'} \,.$$

Like with Lemma 6.4.3, we could try and show that the probability of a particular likelihood exponent is equal to the hitting probability of the union of certain bottom SCCs of $\mathcal{D}$. To compute $\mathcal{D}$, we need to compute $\int_{\delta(S,x)=S'} \Psi(x)_{q,q'}$ for each $S, S' \subseteq Q$.

Recall the set of profiles Span-$\mathcal{I}(\Gamma_m)$ from Example 5.1.1. If $\Psi_{q,q'}$ is an interval-bounded monomial function for each $q, q' \in Q$ then the transition probabilities of $\mathcal{D}$ can be computed using standard integration rules for polynomials.

However, assume $\Psi_{q,q'}(x) = A_{q,q'}\lambda_{q,q'} \exp(-\lambda_{q,q'}x)\chi_{(a_{q,q'},b_{q,q'})}$ for all $x \in (0,\infty)$ where $\lambda_{q,q'}, A_{q,q'} \in \mathbb{Q}_+$ and $(a_{q,q'}, b_{q,q'}) \in \mathbb{I}$ for all $q, q' \in Q$. Then, for $S, S' \subseteq Q$ we have that $\{x \in \mathbb{R} \mid \delta(S,x) = S'\}$ is a union of disjoint open intervals in $\mathbb{I}$. Collect all the end points of these intervals into a set $\{c_1, \ldots, c_K\}$. Then, for $k \in [K]$, $q, q' \in Q$ we may compute $B_{k,q,q'} \in \mathbb{Q}$ such that $\int_{\delta(S,x)=S'} \Psi(x)_{q,q'} = \sum_{i=1}^{K} \sum_{q,q' \in Q} B_{i,q,q'} e^{c_k \lambda_{q,q'}}$. Thus, $\mathcal{D}$ is symbolically computable as a linear combination of rational powers of the number e.

In some cases, we can compute the Markov chain $\mathcal{D}$. In other cases, we can only compute it symbolically. Therefore, we might define two profile languages with different versions of Theorem 6.4.1 for each.

Appendices

A
Missing Proofs

A.1 Proof of Lemma 3.5.2

Lemma 3.5.2. *Suppose $f_1, \dots, f_n \in \mathcal{L}_1(\Sigma, \lambda)$. Then, the f_i are linearly dependent if and only if all alternant matrices for the f_i are singular.*

Proof. Suppose that the $f_1, \dots, f_n$ are linearly dependent. Then, there exist $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ that are not all 0 such that $\sum_{i=1}^{n} \lambda_i f_i(x) = 0$ for all $x \in \Sigma$. The same dependence holds for the columns of any alternant matrix for the f_i. This proves the forward implication.

Write $M_{f_1,\dots,f_n}(x_1, \dots, x_n)$ for the alternant matrix generated by the functions $f_1, \dots, f_n$ and input points $x_1, \dots, x_n$ and let $G_{f_1,\dots,f_n} : \Sigma^n \to \mathbb{R}$ be given by $G_{f_1,\dots,f_n}(x_1, \dots, x_n) = \det M_{f_1,\dots,f_n}(x_1, \dots, x_n)$.

For the converse implication, it suffices to show that $G_{f_1,\dots,f_n} = 0$ on Σ^n implies that $\{f_1, \dots, f_n\}$ is linearly dependent. We proceed by induction on the number n of functions. Suppose $n = 1$ with single function f. If for all $x \in \Sigma$, $0 = G_f(x) = f(x)$ then clearly $f = 0$ and $\{0\}$ is a linearly dependent set in any vector space.

Now suppose that for $n \geq 1$ and arbitrary functions $g_1, \dots, g_n : \Sigma \to \mathbb{R}$, $G_{g_1,\dots,g_n} = 0$ implies that $g_1, \dots, g_n$ are linearly dependent. Let $f_1, \dots, f_{n+1} : \Sigma \to \mathbb{R}$ and $x_1, \dots, x_{n+1}$. A Laplace expansion of $G_{f_1,\dots,f_{n+1}}(x_1, \dots, x_{n+1})$ along the first row of $M_{f_1,\ldots,f_{n+1}}(x_1,\ldots,x_{n+1})$ gives

$$\begin{aligned} G_{f_1,\ldots,f_{n+1}}(x_1,\ldots,x_{n+1}) &= f_1(x_1)G_{f_2,\ldots,f_{n+1}}(x_2,\ldots,x_{n+1}) \\ &\quad + \cdots \\ &\quad + (-1)^n f_{n+1}(x_1)G_{f_1,\ldots,f_n}(x_2,\ldots,x_{n+1}). \end{aligned}$$

Suppose $G_{f_1,\ldots,f_{n+1}}(x_1,\ldots,x_{n+1}) = 0$ holds for all $x_1,\ldots,x_n$. We distinguish between two cases.

- Either there exist $x_2,\ldots,x_{n+1} \in \Sigma$ such that the cofactors

  $$G_{f_2,\ldots,f_{n+1}}(x_2,\ldots,x_{n+1}), G_{f_1,f_3,\ldots,f_{n+1}}(x_2,\ldots,x_{n+1}),\ldots, G_{f_1,\ldots,f_n}(x_2,\ldots,x_{n+1})$$

  are not all 0. This establishes a linear dependence in $f_1,\ldots,f_{n+1}$.

- Or all cofactors are 0 for all $x_2,\ldots,x_{n+1}$. Then, in particular, $G_{f_2,\ldots,f_{n+1}}(x_2,\ldots,x_{n+1}) = 0$ for all $x_2,\ldots,x_{n+1}$. By the induction hypothesis it follows that the functions $f_2,\ldots,f_{n+1}$ are linearly dependent. Hence, so are $f_1,\ldots,f_{n+1}$.

In either case it follows that $f_1,\ldots,f_{n+1}$ are linearly dependent. □

A.2 Equivalent Conditions in the Distinguishability Algorithm

We have the following lemma used in the distinguishability algorithm for finite-observation HMMs.

Lemma A.2.1. *Let (Q,Σ,M) be a finite-observation HMM and let $\mathcal{B}'$ be a basis for*

$$V := \operatorname{span}\left\{ \begin{pmatrix} M(w) & 0 \\ 0 & -M(w) \end{pmatrix} \mathbb{1}^T \;\middle|\; w \in \Sigma^* \right\}.$$

Let $\mathcal{B}$ be a basis for $\{M(w)\mathbb{1}^T \mid w \in \Sigma^*\}$ and let $\mu_1,\mu_2 \in \mathbb{Q}^{|Q|}$ be row vectors. Then $(\mu_1\ \mu_2)\cdot b' = 0$ for all $b' \in \mathcal{B}'$ if and only if $\mu_1 - \mu_2 \cdot b = 0$ for all $b \in \mathcal{B}$. Here, $(\mu_1\ \mu_2) \in [0,1]^{2|Q|}$ is the row vector obtained by gluing μ_1,μ_2 together. Note that $(\mu_1\ \mu_2)\cdot b$ is a scalar.

Proof. We have

$$
\begin{aligned}
& (\mu_1\ \mu_2)\cdot b' = 0 \quad \forall\, b' \in \mathcal{B}' \\
\iff & (\mu_1\ \mu_2) \begin{pmatrix} M(w) & 0 \\ 0 & -M(w) \end{pmatrix} \mathbb{1}^T = 0 \quad \forall w \in \Sigma^* \\
\iff & \mu_1 M(w) - \mu_2 M(w) = 0 \quad \forall w \in \Sigma^* \\
\iff & (\mu_1 - \mu_2) M(w) = 0 \quad \forall w \in \Sigma^* \\
\iff & (\mu_1 - \mu_2) \perp \operatorname{span}\{M(w)\mathbb{1}^T \mid w \in \Sigma^*\} \\
\iff & (\mu_1 - \mu_2)\cdot b = 0 \quad \forall\, b \in \mathcal{B}.
\end{aligned}
$$

□

A.3 Proofs Concerning Analytic Functions

Recall the identity theorem and related corollary:

Theorem 5.2.1. *[Identity Theorem] Let $f : \mathbb{R} \to \mathbb{R}$ be analytic on $\mathbb{R}$. If $f(x) = 0$ for all $x \in S \subseteq \mathbb{R}$, where S has an accumulation point, then $f(x) = 0$ for all $x \in \mathbb{R}$.*

Corollary 5.2.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be analytic on $\mathbb{R}$. If $f(x) = 0$ for all $x \in S \subseteq \mathbb{R}$, where $\lambda_{Leb}(S) > 0$, then $f(x) = 0$ for all $x \in \mathbb{R}$.*

We prove Corollary 5.2.2 using the following lemma

Lemma A.3.1. *Suppose $S \subseteq \mathbb{R}$ has $\lambda_{Leb}(S) > 0$ then S has an accumulation point.*

Proof. We prove the contrapositive. Suppose S has no accumulation points. Then for all $x \in S$ there is $\varepsilon_x > 0$ such that $B(x, \varepsilon_x) \cap S = \{x\}$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$ there is a rational $q_x \in B(x, \varepsilon_x)$ Hence $x = q_x$. It follows that S contains only rationals and hence is countable. Therefore, $\lambda_{Leb}(S) = 0$. □

We may now also prove the following from the main body.

Lemma 5.1.2. *Let $D \subseteq \mathbb{R}$ and let $f : D \to \mathbb{R}$ be analytic on D. If an analytic continuation of f to $\mathbb{R}$ exists, then it is unique.*

Proof. Suppose $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ are analytic continuations of f to $\mathbb{R}$. Then, $f_1(x) - f_2(x) = 0$ for all $x \in D$. D is an open set and so $\lambda(D) > 0$. Therefore, by Corollary 5.2.2 we have $f_1(x) = f_2(x)$ for all $x \in \mathbb{R}$. □

The following two lemmas are used in Lemma A.3.4 to establish that some common pdfs are analytic on $\mathbb{R}$.

Lemma A.3.2. *Define $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = \sum_{n=0}^{N} b_n x^n$ for some $b_0, \ldots, b_N \in \mathbb{R}$. Then, f is analytic on $\mathbb{R}$.*

Proof. Let $x_0 \in \mathbb{R}$. For $x \in \mathbb{R}$ we have $f(x + x_0) = \sum_{n=0}^{N} b_n (x + x_0)^n = \sum_{n=0}^{N} a_n x^n$ where $a_0, \ldots, a_N$ depend on x_0 and $b_0, \ldots, b_N$ and are obtained by expanding each bracket and collecting coefficients with the same power of x. Thus, $f(x) = \sum_{n=0}^{N} a_n (x - x_0)^n$ and so f is analytic on $\mathbb{R}$. □

Lemma A.3.3. *Let $A, B \in \mathbb{R}$ and define $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = Ae^{Bx}$. Then, f is analytic on $\mathbb{R}$.*

Proof. Let $x_0 \in \mathbb{R}$. For $x \in \mathbb{R}$ we have

$$f(x + x_0) = Ae^{Bx + Bx_0} = Ae^{Bx_0} e^{Bx} = Ae^{Bx_0} \sum_{n=0}^{\infty} B^n \frac{x^n}{n!}.$$

It follows that $f(x) = \sum_{n=0}^{\infty} \frac{Ae^{Bx_0} B^n}{n!} (x - x_0)^n$ and so f is analytic by setting $a_n = \frac{Ae^{Bx_0} B^n}{n!}$. □

Lemma A.3.4. *Let $\nu \in \mathbb{Q}_+$, $k \in \mathbb{N} \setminus \{1\}$, $\lambda \in \mathbb{Q}_+$, $a \in \mathbb{Q}_+$, $\mu \in \mathbb{Q}$ and $\sigma \in \mathbb{Q}_+$. The functions $f_t, f_\chi, f_e, f_{mb}, f_g : \mathbb{R} \to \mathbb{R}$ are analytic on $\mathbb{R}$ where*

1. $f_t(x) = \frac{\tilde{\Gamma}((\nu+1)2)}{\sqrt{\nu\pi}\tilde{\Gamma}(\nu/2)} (1 + \frac{x^2}{\nu})^{-(\nu+1)/2}$
2. $f_\chi(x) = \frac{1}{2^{k/2}\tilde{\Gamma}(k/2)} x^{k/2-1} \exp(-x/2)$
3. $f_e(x) = \lambda \exp(-\lambda x)$
4. $f_{mb} = \sqrt{\frac{2}{\pi}} \frac{1}{a^3} x^2 \exp(-x^2/(2a^2))$
5. $f_g = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

Proof of Lemma A.3.4. f_t is a polynomial so is analytic by Lemma A.3.2. f_e is analytic on $\mathbb{R}$ by Lemma A.3.3. Consider $f, g : \mathbb{R} \to \mathbb{R}$ where $f(x) = x^{k/2-1}$ and $g(x) = \exp(-x/2)$. Then, f and g are analytic on $\mathbb{R}$ by Lemma A.3.2 and Lemma A.3.3 respectively. The product of two analytic functions is analytic by Proposition 1.1.4 from [34] and so f_χ is analytic on $\mathbb{R}$. A similar argument shows that f_{mb} is analytic on $\mathbb{R}$. We showed f_g is analytic in Example 5.1.3. $\square$

References

[1] L.R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[2] F.-S. Chen, C.-M. Fu, and C.-L. Huang. “Hand gesture recognition using a real-time tracking method and hidden Markov models”. In: *Image and Vision Computing* 21.8 (2003), pp. 745–758.

[3] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. “Wavelet-based statistical signal processing using hidden Markov models”. In: *IEEE Transactions on Signal Processing* 46.4 (Apr. 1998), pp. 886–902.

[4] P. Ailliot, C. Thompson, and P. Thomson. “Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions”. In: *Journal of the Royal Statistical Society* 58.3 (2009), pp. 405–426.

[5] S.R. Eddy. “What is a hidden Markov model?” In: *Nature Biotechnology* 22.10 (Oct. 2004), pp. 1315–1316.

[6] G.A. Churchill. “Stochastic models for heterogeneous DNA sequences”. In: *Bulletin of Mathematical Biology* 51.1 (1989), pp. 79–94.

[7] R. Durbin. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[8] A. Krogh, B. Larsson, G. von Heijne, and E.L.L. Sonnhammer. “Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes”. In: *Journal of Molecular Biology* 305.3 (2001), pp. 567–580.

[9] M. Alexandersson, S. Cawley, and L. Pachter. “SLAM: Cross-Species Gene Finding and Alignment with a Generalized Pair Hidden Markov Model”. In: *Genome Research* 13 (2003), pp. 469–502.

[10] M. Kwiatkowska, G. Norman, and D. Parker. “PRISM 4.0: Verification of Probabilistic Real-time Systems”. In: *Proceedings of Computer Aided Verification (CAV).* Vol. 6806. LNCS. Springer, 2011, pp. 585–591.

[11] C. Dehnert, S. Junges, J.-P. Katoen, and M. Volk. “A Storm is Coming: A Modern Probabilistic Model Checker”. In: *Proceedings of Computer Aided Verification (CAV).* Springer, 2017, pp. 592–600.

[12] M.P. Schützenberger. “On the definition of a family of automata”. In: *Information and Control* 4.2 (1961), pp. 245–270.

[13] Azaria Paz. *Introduction to Probabilistic Automata (Computer Science and Applied Mathematics).* Orlando, FL, USA: Academic Press, Inc., 1971.

[14] Wen-Guey Tzeng. “A Polynomial-time Algorithm for the Equivalence of Probabilistic Automata”. In: *SIAM J. Comput.* 21.2 (Apr. 1992), pp. 216–227. URL: `http://dx.doi.org/10.1137/0221017`.

[15] C. Cortes, M. Mohri, and A. Rastogi. “L_p Distance and Equivalence of Probabilistic Automata”. In: *International Journal of Foundations of Computer Science* 18.04 (2007), pp. 761–779.

[16] M.S. Alvim, M.E. Andrés, C. Palamidessi, and P. van Rossum. “Safe Equivalences for Security Properties”. In: *Theoretical Computer Science*. Springer, 2010, pp. 55–70.

[17] S. Kiefer, A.S. Murawski, J. Ouaknine, B. Wachter, and J. Worrell. “Language Equivalence for Probabilistic Automata”. In: *Proceedings of the 23rd International Conference on Computer Aided Verification (CAV)*. Vol. 6806. LNCS. Springer, 2011, pp. 526–540.

[18] M.S. Bauer, R. Chadha, and M. Viswanathan. “Modular Verification of Protocol Equivalence in the Presence of Randomness”. In: *Computer Security – ESORICS 2017*. Springer, 2017, pp. 187–205.

[19] C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen. “Model-checking algorithms for continuous-time Markov chains”. In: *IEEE Transactions on Software Engineering* 29.6 (2003), pp. 524–541.

[20] T. Chen, M. Diciolla, M.Z. Kwiatkowska, and A. Mereacre. “Time-Bounded Verification of CTMCs against Real-Time Specifications”. In: *Proceedings of Formal Modeling and Analysis of Timed Systems (FORMATS)*. Vol. 6919. LNCS. Springer, 2011, pp. 26–42.

[21] W.K. Grassmann. “Finding transient solutions in Markovian event systems through randomization”. In: *Numerical solution of Markov chains*. 1991, pp. 357–371.

[22] H. Hermanns, J. Krčál, and J. Křetínský. “Probabilistic Bisimulation: Naturally on Distributions”. In: *CONCUR 2014 – Concurrency Theory*. Springer, 2014, pp. 249–265.

[23] Taolue Chen and Stefan Kiefer. “On the Total Variation Distance of Labelled Markov Chains”. In: *Proceedings of the Joint Meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic (CSL) and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*. Vienna, Austria, 2014, 33:1–33:10.

[24] Stefan Kiefer and A. Prasad Sistla. “Distinguishing Hidden Markov Chains”. In: *Proceedings of the 31st Annual Symposium on Logic in Computer Science (LICS)*. New York, USA: ACM, 2016, pp. 66–75.

[25] N. Bertrand, S. Haddad, and E. Lefaucheux. “Accurate Approximate Diagnosability of Stochastic Systems”. In: *Proceedings of Language and Automata Theory and Applications (LATA)*. Vol. 9618. Lecture Notes in Computer Science. Springer, 2016, pp. 549–561. URL: https://doi.org/10.1007/978-3-319-30000-9%5C_42.

[26] S. Akshay, H. Bazille, E. Fabre, and B. Genest. “Classification Among Hidden Markov Models”. In: *Proceedings of the Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*. Vol. 150. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019, 29:1–29:14. URL: https://doi.org/10.4230/LIPIcs.FSTTCS.2019.29.

[27] A. Wald. “Sequential Tests of Statistical Hypotheses”. In: *The Annals of Mathematical Statistics* 16.2 (1945), pp. 117–186. URL: https://doi.org/10.1214/aoms/1177731118.

[28] A. Wald and J. Wolfowitz. “Optimum Character of the Sequential Probability Ratio Test”. In: *The Annals of Mathematical Statistics* 19.3 (1948), pp. 326–339. URL: `http://www.jstor.org/stable/2235638` (visited on 04/21/2022).

[29] V.Yu. Protasov and R.M. Jungers. “Lower and upper bounds for the largest Lyapunov exponent of matrices”. In: *Linear Algebra and its Applications* 438.11 (2013), pp. 4448–4468. URL: `https://www.sciencedirect.com/science/article/pii/S002437951300089X`.

[30] D. Sutter, O. Fawzi, and R. Renner. “Bounds on Lyapunov Exponents via Entropy Accumulation”. In: *IEEE Transactions on Information Theory* 67.1 (2021), pp. 10–24.

[31] Oscar Darwin and Stefan Kiefer. “Equivalence of Hidden Markov Models with Continuous Observations”. In: *40th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2020)*. Ed. by Nitin Saxena and Sunil Simon. Vol. 182. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020, 43:1–43:14. URL: `https://drops.dagstuhl.de/opus/volltexte/2020/13284`.

[32] Oscar Darwin and Stefan Kiefer. “On the Sequential Probability Ratio Test in Hidden Markov Models”. In: *33rd International Conference on Concurrency Theory (CONCUR 2022)*. Ed. by Bartek Klin, Sławomir Lasota, and Anca Muscholl. Vol. 243. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022, 9:1–9:16. URL: `https://drops.dagstuhl.de/opus/volltexte/2022/17072`.

[33] B.A. Braun, S. Jana, and D. Boneh. *Robust and Efficient Elimination of Cache and Timing Side Channels*. 2015. arXiv: `1506.00189 [cs.CR]`.

[34] Steven G. Krantz and Harold R. Parks. “A Primer of Real Analytic Functions”. In: 1992.

[35] Maxime Bôcher. “Certain cases in which the vanishing of the Wronskian is a sufficient condition for linear dependence”. In: *Transactions of the American Mathematical Society* 2 (1901), pp. 139–149.

[36] E. Landau. *Handbuch der lehre von der verteilung der primzahlen*. Handbuch der lehre von der verteilung der primzahlen Bd. 1. B. G. Teubner, 1909. URL: `https://books.google.de/books?id=81m4AAAAIAAJ`.

[37] R. Langrock, B. Swihart, B. Caffo, N. Punjabi, and C. Crainiceanu. “Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms”. In: *Statistics in medicine* 32 (Aug. 2013).

[38] C.M. Papadimitriou. *Computational complexity*. Addison-Wesley, 1994.

[39] A. Borodin. “On Relating Time and Space to Size and Depth”. In: *SIAM Journal of Computing* 6.4 (1977), pp. 733–744. URL: `https://doi.org/10.1137/0206054`.

[40] Jui-Yi Kao, Narad Rampersad, and Jeffrey Shallit. “On NFAs where all states are final, initial, or both”. In: *Theoretical Computer Science* 410.47 (2009), pp. 5010–5021. URL: `https://www.sciencedirect.com/science/article/pii/S0304397509005477`.

[41] A. Borodin, J. von zur Gathen, and J.E. Hopcroft. "Fast Parallel Matrix and GCD Computations". In: *Information and Control* 52.3 (1982), pp. 241–256. URL: https://doi.org/10.1016/S0019-9958(82)90766-5.

[42] Vladimir Bogachev. *Measure Theory*. Vol. 1. Jan. 2007, pp. 1–575.

[43] Theodore J. Sheskin. "Conditional mean first passage time in a Markov chain". In: *International Journal of Management Science and Engineering Management* 8.1 (2013), pp. 32–37.

[44] K. Etessami, A. Stewart, and M. Yannakakis. "A Note on the Complexity of Comparing Succinctly Represented Integers, with an Application to Maximum Probability Parsing". In: *ACM Trans. Comput. Theory* 6.2 (2014), 9:1–9:23. URL: https://doi.org/10.1145/2601327.

[45] V.Yu. Protasov. "Asymptotics of products of nonnegative random matrices". In: *Functional Analysis and Its Applications* 47 (2013), pp. 138–147.

[46] H. Furstenberg and H. Kesten. "Products of Random Matrices". In: *The Annals of Mathematical Statistics* 31.2 (1960), pp. 457–469. URL: http://www.jstor.org/stable/2237962 (visited on 03/25/2023).

[47] László Gerencsér, G. Michaletzky, and Zsanett Orlovits. "Stability of block-triangular stationary random matrices". In: *Systems & Control Letters* (Aug. 2008), pp. 620–625.

[48] B.-H. Juang and L. R. Rabiner. "A Probabilistic Distance Measure for Hidden Markov Models". In: *AT&T Technical Journal* 64.2 (1985), pp. 391–408. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1985.tb00439.x.

[49] B. Chen and P. Willett. "Detection of hidden Markov model transient signals". In: *IEEE Transactions on Aerospace and Electronic Systems* 36.4 (2000), pp. 1253–1268.

[50] C.-D. Fuh. "SPRT and CUSUM in hidden Markov models". In: *The Annals of Statistics* 31.3 (2003), pp. 942–977. URL: https://doi.org/10.1214/aos/1056562468.

[51] E. Grossi and M. Lops. "Sequential Detection of Markov Targets With Trajectory Estimation". In: *IEEE Transactions on Information Theory* 54.9 (2008), pp. 4144–4154.