On the Sequential Probability Ratio Test in Hidden Markov Models

Oscar Darwin
Department of Computer Science, Oxford University, UK

Stefan Kiefer
Department of Computer Science, Oxford University, UK

Abstract

We consider the Sequential Probability Ratio Test applied to Hidden Markov Models. Given two 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 systems. Further, we give complexity results about the execution time taken by the Sequential Probability Ratio Test.

2012 ACM Subject Classification Theory of computation → Random walks and Markov chains; Mathematics of computing → Stochastic processes; Theory of computation → Logic and verification

Keywords and phrases Markov chains, hidden Markov models, probabilistic systems, verification

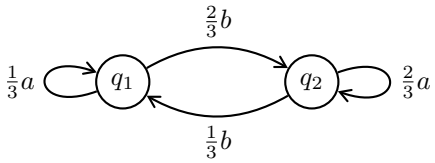
Digital Object Identifier 10.4230/LIPIcs.CONCUR.2022.9

Related Version *Full Version*: `https://arxiv.org/abs/2207.14088`

Acknowledgements The authors thank anonymous referees for valuable suggestions.

1 Introduction

A (discrete-time, finite-state) *Hidden Markov Model (HMM)* (often called *labelled Markov chain*) has a finite set Q of states and for each state a probability distribution over its possible successor states. Every state is associated with a probability transition over a successor state and an emitted letter (*observation*). For example, consider the following HMM:



In state q_1, the probability of emitting a and the next state being also q_1 is $\frac{1}{3}$, and the probability of emitting b and the next state being q_2 is $\frac{2}{3}$. An HMM is typically viewed as a producer of a finite or infinite word of emitted observations. For example, starting in q_1, the probability of producing a word with prefix aba is $\frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3}$, whereas starting in q_2, the probability of aba is $\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3}$. The random sequence of states is considered not observable (which explains the term *hidden* in HMM).

HMMs are widely employed in fields such as speech recognition (see [28] for a tutorial), gesture recognition [6], signal processing [10], and climate modeling [1]. HMMs are heavily used in computational biology [14], more specifically in DNA modeling [8] and biological sequence analysis [13], including protein structure prediction [22] and gene finding [3]. In computer-aided verification, HMMs are the most fundamental model for probabilistic systems; model-checking tools such as Prism [23] or Storm [12] are based on analyzing HMMs efficiently.

One of the most fundamental questions about HMMs is whether two initial distributions are *(trace) equivalent*, i.e., generate the same distribution on infinite observation sequences. In the example above, we argued that (the Dirac distributions on) the states q_1, q_2 are not equivalent. The equivalence problem is very well studied and can be solved in polynomial time using algorithms that are based on linear algebra [29, 25, 31, 9]. The equivalence problem has applications in verification, e.g., of randomised anonymity protocols [20].

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 [7]. There it was shown that 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 [21] and diagnosability [4, 2] of stochastic systems.

Distributions π_1, π_2 that are distinguishable (i.e., $d(\pi_1, \pi_2) = 1$) can nevertheless be "hard" to distinguish. In our example above, (the Dirac distributions on) q_1, q_2 are distinguishable. 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.

To make this more precise, 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 example above, we argued that $\mathbb{P}_{q_1}(aba\Sigma^\omega) = \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3}$ 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., [7, 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 increase or decrease of $\ln L_n$ as a measure of *how* distinguishable two distinguishable distributions π_1, π are.

The log-likelihood ratio plays a central role in the *sequential probability ratio test (SPRT)* [32], which is optimal [33] among sequential hypothesis tests (such tests attempt to decide between two hypotheses without fixing the sample size in advance). In terms of an HMM and two initial distributions π_1, π_2, the SPRT attempts to 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.

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. As the most fundamental contribution of this paper, we show that the average slope exists almost surely and that any HMM with m states has at most $m^2 + 1$ likelihood exponents.

The rest of the paper is organised as follows. In Section 3 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. This connection motivates our results on likelihood exponents in the rest of the paper. In Section 4 we prove complexity results concerning the probability that the average slope equals a particular likelihood exponent. In Section 5 we show that the average slope exists almost surely and prove our bound on the number of likelihood exponents. Further, we show that the likelihood exponents can be efficiently expressed in terms of Lyapunov exponents. In Section 6 we show that for *deterministic* HMMs one can compute likelihood exponents in polynomial time. We conclude in Section 7.

2 Preliminaries

We write $\mathbb{N}$ for the set of non-negative integers. For $d \in \mathbb{N}$ we write $[d] = \{1, \ldots, d\}$. For a finite set Q, vectors $\mu \in \mathbb{R}^Q$ are viewed as row vectors, and their transpose (a column vector) is denoted by $\mu^\top$. The norm $\|\mu\|$ is assumed to be the l_1 norm: $\|\mu\| = \sum_{q\in Q} |\mu_q|$. We write $\vec{0}, \vec{1}$ for the vectors all whose entries are 0, 1, respectively. 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 $M \in [0,1]^{Q\times Q}$ is *stochastic* if $\vec{1}^\top = M\vec{1}^\top$. We often identify vectors $\mu \in [0,1]^Q$ such that $\|\mu\| = 1$ with the corresponding probability distribution on Q. For $\mu \in [0,\infty)^Q$ we write $\mathrm{supp}(\mu) := \{q \in Q \mid \mu_q > 0\}$.

For a finite alphabet Σ and $n \in \mathbb{N}$ we denote by $\Sigma^n, \Sigma^*, \Sigma^\omega$ the sets of length-n words, finite words, infinite words, respectively. For $w \in \Sigma^\omega$ we write w_n for the length-n prefix of w.

A *Hidden Markov Model* (HMM) is a triple $\mathcal{H} = (Q, \Sigma, \Psi)$ where Q is a finite set of states, Σ is a set of observations (or "letters"), and the function $\Psi : \Sigma \to [0,1]^{Q\times Q}$ specifies the transitions such that $\sum_{a\in\Sigma} \Psi(a)$ is stochastic. 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. The *embedded* Markov chain of an HMM (Q, Σ, Ψ) is the Markov chain $(Q, \sum_{a\in\Sigma} \Psi(a))$. We say that an HMM is *strongly connected* if the graph of its embedded Markov chain is.

▶ **Example 1.** The HMM from the introduction is the triple $\mathcal{H} = (\{q_1, q_2\}, \{a, b\}, \Psi)$ with $\Psi(a) = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{2}{3} \end{pmatrix}$ and $\Psi(b) = \begin{pmatrix} 0 & \frac{2}{3} \\ \frac{1}{3} & 0 \end{pmatrix}$. The embedded Markov chain is $(\{q_1, q_2\}, \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix})$.

Fix an 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. We call a finite sequence $v = q_0 a_1 q_1 \cdots a_n q_n \in Q(\Sigma Q)^*$ a *path* and $v(\Sigma Q)^\omega$ a *cylinder set* and an infinite sequence $q_0 a_1 q_1 a_2 q_2 \cdots \in Q(\Sigma Q)^\omega$ a *run*. To $\mathcal{H}$ and an *initial probability distribution* $\pi \in [0,1]^Q$ we associate the probability space $(Q(\Sigma Q)^\omega, \mathcal{G}^*, \mathbb{P}_\pi)$ where $\mathcal{G}^*$ is the σ-algebra generated by the cylinder sets and $\mathbb{P}_\pi$ is the unique probability measure with $\mathbb{P}_\pi(q_0 a_1 q_1 \cdots a_n q_n (\Sigma Q)^\omega) = \pi_q \prod_{i=1}^n \Psi(a_i)_{q_{i-1},q_i}$. As the states are often irrelevant, for $E \subseteq \Sigma^\omega$ and $E\uparrow := \{q_0 a_1 q_1 a_2 q_2 \cdots \mid a_1 a_2 \cdots \in E\} \in \mathcal{G}^*$ we view also E as an event and may write $\mathbb{P}_\pi(E)$ to mean $\mathbb{P}_\pi(E\uparrow)$. In particular, for $w \in \Sigma^*$ we

have $\mathbb{P}_\pi(w\Sigma^\omega) = \|\pi\Psi(w)\|$. For $E \subseteq \Sigma^\omega$ we write $\mathbb{1}_E$ for the indicator random variable with $\mathbb{1}_E(w) = 1$ if $w \in E$ and $\mathbb{1}_E(w) = 0$ if $w \notin E$. 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 $\mathbb{E}_q$.

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.

Let (Q, Σ, Ψ) be an HMM and let π_1, π_2 be two initial distributions. The *total variation distance* is $d(\pi_1, \pi_2) := \sup_{E^\uparrow \in \mathcal{G}^*} |\mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)|$. This supremum is actually a maximum due to Hahn's decomposition theorem; i.e., there is an event $E \subseteq \Sigma^\omega$ such that $d(\pi_1, \pi_2) = \mathbb{P}_{\pi_1}(E) - \mathbb{P}_{\pi_2}(E)$. We call π_1 and π_2 *distinguishable* if $d(\pi_1, \pi_2) = 1$. Distinguishability is decidable in polynomial time [7].

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 $L_n(w) = \frac{\|\pi_1\Psi(w_n)\|}{\|\pi_2\Psi(w_n)\|}$. Based on results from [7] we have the following lemma.

▶ **Lemma 2.** *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.*

▶ **Example 3.** We illustrate convergence of the likelihood ratio using an example from [24] 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. We describe in [11] how we obtained from this an HMM $\mathcal{H} = (Q, \Sigma, \Psi)$ with (finite) observation alphabet $\Sigma = \{a_1, \ldots, a_5\}$ and two initial distributions π_1, π_2 corresponding to healthy and diseased individuals, respectively. Using the algorithm from [7] one can show that π_1 and π_2 are distinguishable.

We sampled runs of $\mathcal{H}$ 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 1.

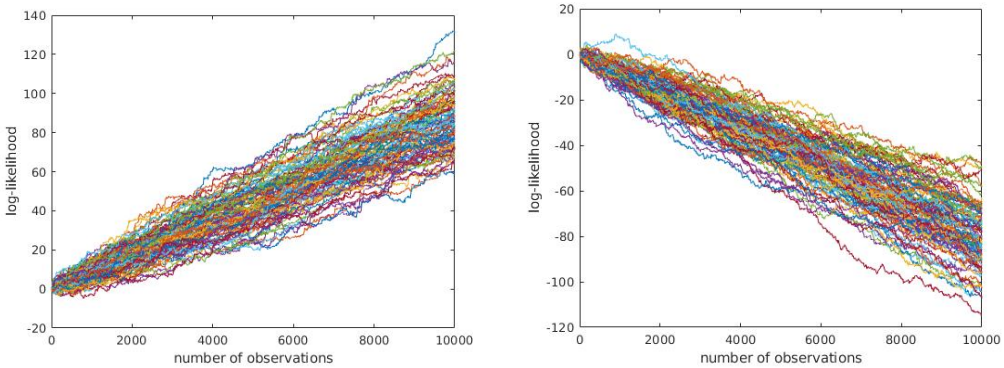


Figure 1 The two images show two log-likelihood plots of sample runs produced by π_1 and π_2, respectively.

By Lemma 2.2 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 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).

3 Sequential Probability Ratio Test

Fix an HMM $H = (Q, \Sigma, \Psi)$ for the rest of the paper. 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 correctness property.

▶ **Proposition 4.** *Suppose π_1 and π_2 are distinguishable. Let $\alpha, \beta \in (0,1)$. By choosing $A = \ln \frac{\alpha}{1-\beta}$ and $B = \ln \frac{1-\alpha}{\beta}$, we have $\mathbb{P}_{\pi_1}(\mathrm{SPRT}_{\alpha,\beta} = \pi_2) \leq \alpha$ and $\mathbb{P}_{\pi_2}(\mathrm{SPRT}_{\alpha,\beta} = \pi_1) \leq \beta$.*

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 $\overline{\mathrm{SPRT}}_{\beta,\alpha} = \mathrm{SPRT}_{\alpha,\beta}$. 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 Lemma 2.2.

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

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

(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 [32] shows that the expected stopping time $\mathbb{E}_{s_2}[N_{\alpha,\beta}]$ and ℓ are inversely proportional:

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

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. [21, Example 6].

[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 $i = 1, 2$.

Nevertheless, in Figure 1 we observed that $\ln L_n$ appears to decrease linearly (on the π_2 plot). Indeed, we show in Theorem 8 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:

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

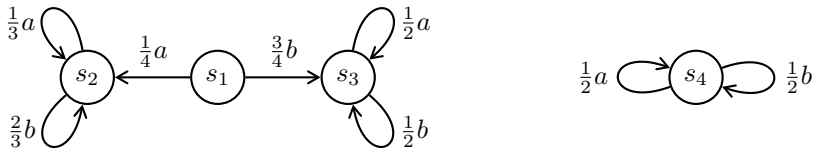
The number ℓ is called a likelihood exponent, as defined generally in the following definition.

▶ **Definition 5.** *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 Lemma 2.1 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 $\mathcal{H}$. For $\ell \in \Lambda$ we define the event $E_\ell = \{\lim_{n\to\infty} \frac{1}{n} \ln L_n = \ell\}$.

▶ **Example 6.** In the case of Example 3 we have $\Lambda_{\pi_1,\pi_2} = \{\ell\}$ where the slope of the right hand side of Figure 1 suggests that $\ell \approx -\frac{80}{10000} = -0.008$.

▶ **Example 7.** 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. 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 paper.

▶ **Theorem 8.** *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 $|\Lambda| \leq |Q|^2 + 1$.*

It follows from a stronger theorem, Theorem 23, which we prove in Section 5.

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 1) to lead to faster termination. In the two-state case, Wald's formula (1) becomes

$$\mathbb{E}_{s_2}[N_{\alpha,\beta}] = \frac{\beta \ln \frac{1-\alpha}{\beta} + (1-\beta) \ln \frac{\alpha}{1-\beta}}{\ell} \sim \frac{\ln \alpha}{\ell} \text{ (as } \alpha, \beta \to 0), \tag{2}$$

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 9 below we generalise Equation (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 9** (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 $\mathbb{E}_{\pi_2}\left[N_{\alpha,\beta} \mid E_\ell\right] \sim \dfrac{\ln \alpha}{\ell}$ (as $\alpha, \beta \to 0$).*
2. *If $\ell = 0$ then there exist $\alpha, \beta > 0$ such that $\mathbb{E}_{\pi_2}\left[N_{\alpha,\beta} \mid E_\ell\right] = \infty$.*
3. *If $\ell = -\infty$ then $\sup_{\alpha,\beta} \mathbb{E}_{\pi_2}\left[N_{\alpha,\beta} \mid E_\ell\right] < \infty$.*

The theorem above 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.

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

3.2.1 Likelihood Exponent 0

▶ **Example 10.** We continue with Example 7 to illustrate the second case in Theorem 9. 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 9 says that the expectation of $N_{\alpha,\beta}$ conditioned under E_0 is infinite. The following proposition 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 $\alpha \le \alpha'$ and $\beta \le \beta'$.

▶ **Proposition 11.** *The following two equalities hold up to $\mathbb{P}_{\pi_2}$-null sets:*

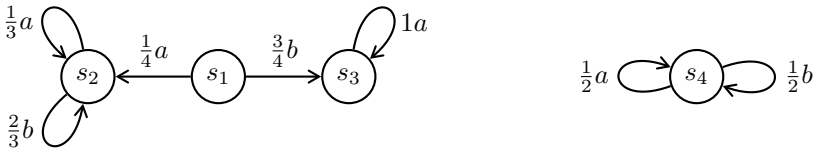
$$E_0 \;=\; \left\{\lim_{n\to\infty} L_n > 0\right\} \;=\; \bigcup_{\alpha,\beta>0} \{N_{\alpha,\beta} = \infty\}\,.$$

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

▶ **Corollary 12** (using Lemma 2.2). *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 $\alpha, \beta > 0$.*

3.2.2 Likelihood Exponent $-\infty$

▶ **Example 13.** Consider now a modification of Example 7 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 right chain producing an observation which the left chain cannot produce, 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 14.** *The events* $E_{-\infty}$ *and* $\{L_n = 0 \text{ for some } n\}$ *are equal. Thus,* $\sup_{\alpha,\beta} N_{\alpha,\beta} = N_\perp$ *and* $\lim_{\alpha,\beta\to 0} \mathbb{P}_{\pi_2}(N_{\alpha,\beta} < \infty) = \mathbb{P}_{\pi_2}(E_{-\infty})$.

Applying this to Example 13, we obtain $\sup_{\alpha,\beta} \mathbb{E}_{\pi_2}\left[N_{\alpha,\beta} \mid E_{-\infty}\right] = 3$.

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

Conditioned on E_ℓ where $\ell \in (-\infty, 0)$, Theorem 9 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 15.** *Let* $\ell \in \Lambda$ *and assume* $\ell \in (-\infty, 0)$. *We have*

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

In fact, we prove the first part of Theorem 9 using Proposition 15. If there were a bound $M \in \mathbb{N}$ such that $\mathbb{P}_{\pi_2}$-a.s. $\frac{N_{\alpha,\beta}}{-\ln \alpha} \leq M$, the first part of Theorem 9 would follow from Proposition 15 by the dominated convergence theorem. However this is not the case in general. Instead we show in [11] that the set of random variables $\{\frac{N_{\alpha,\beta}}{-\ln \alpha} \mid 0 < \alpha, \beta \leq \frac{1}{2}\}$ is uniformly integrable with respect to the measure $\mathbb{P}_{\pi_2}$ and then use Vitali's convergence theorem.

▶ **Example 16.** Recall Example 3, where $\Lambda = \{\ell\}$. Figure 2 demonstrates the asymptotic

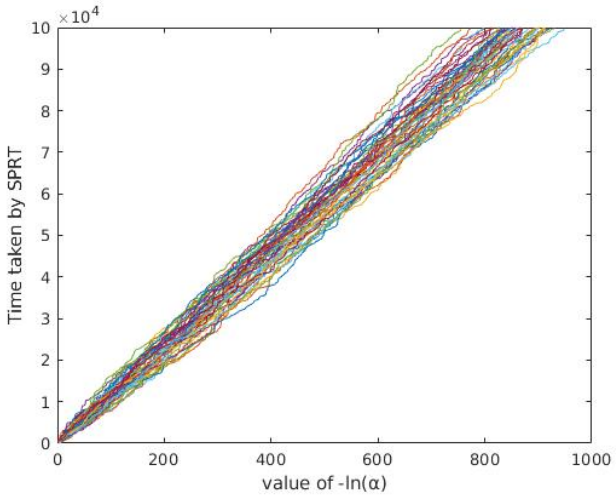


Figure 2 The time taken by the SPRT for $0 \leq -\ln \alpha = -\ln \beta \leq 1000$.

relationship in Proposition 15. Each of the 50 lines correspond to a sample run and we record the value of $N_{\alpha,\beta}$ for $0 \leq -\ln \alpha = -\ln \beta \leq 1000$. From the figure we estimate $-\frac{1}{\ell}$ as $\frac{10^5}{800} = 125$. This coincides with the estimate given in Example 6.

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 rest of the paper.

4 Probability of E_ℓ

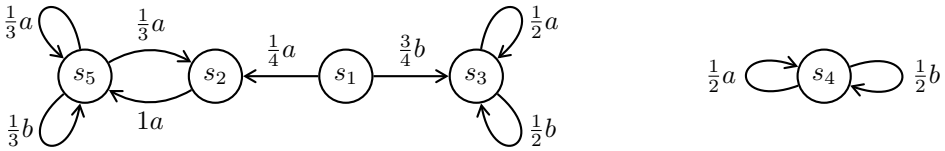
In this section we aim at computing $\mathbb{P}_{\pi_2}(E_\ell)$ for a likelihood exponent ℓ. We show the following theorem.

▶ **Theorem 17.** *Given an HMM and initial distributions π_1, π_2,*

1. *one can compute $\mathbb{P}_{\pi_2}(E_{-\infty})$ and $\mathbb{P}_{\pi_2}(E_0)$ in PSPACE;*
2. *one can decide whether $\mathbb{P}_{\pi_2}(E_0) = 0$ (i.e., $0 \notin \Lambda_{\pi_1,\pi_2}$) in polynomial time;*
3. *deciding whether $\mathbb{P}_{\pi_2}(E_0) = 1$, whether $\mathbb{P}_{\pi_2}(E_{-\infty}) = 0$, and whether $\mathbb{P}_{\pi_2}(E_{-\infty}) = 1$ are all PSPACE-complete problems.*

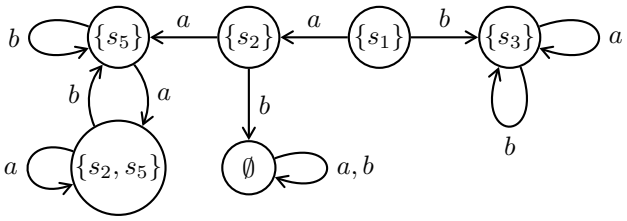
The following example illustrates the construction underlying the PSPACE upper bound.

▶ **Example 18.** Consider another adaption of Example 7.



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 $s_5 \in \mathrm{supp}(e_{s_1}\Psi(aaw))$ 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 14) that there is $\ell \in (-\infty, 0)$ such that $\lim_{n\to\infty} \frac{1}{n} \ln L_n = \ell$. Thus, $\Lambda_{s_1,s_4} = \{-\infty, \ell, 0\}$.

The likelihood ratio L_n is 0 if and only if $\mathrm{supp}(\pi_1\Psi(w_n)) = \emptyset$. In order to track the support of $\pi_1\Psi(w_n)$, 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 an HMM. Therefore, in the following construction, we also keep track of the "current" state of the HMM which produces the observations. For $S \subseteq Q$ and $a \in \Sigma$, define $\delta(S, a) := \{q' \in Q \mid \exists\, q \in S : \Psi(a)_{q,q'} > 0\}$. Define the Markov chain $\mathcal{B} := (2^Q \times Q, T)$ where

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

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 $\pi_1\Psi(w_n)$, 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 19.** *Consider the Markov chain $\mathcal{B} = (2^Q \times Q, T)$ defined above.*

1. *Every bottom SCC of $\mathcal{B}$ 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 Theorem 17.1.

Proof sketch for Theorem 17.1. The Markov chain $\mathcal{B}$ from Lemma 19 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. This PSPACE transducer can also identify the bottom SCCs. 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 Lemma 19.2 and the polynomial-time algorithm for distinguishability from [7]. Finally, to compute $\mathbb{P}_{\pi_2}(E_{-\infty})$ and $\mathbb{P}_{\pi_2}(E_0)$, by Lemma 19.3, 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. Since linear systems of equations can be solved in the complexity class NC, which is included in polylogarithmic space, one can use standard techniques for composing space-bounded transducers to compute $\mathbb{P}_{\pi_2}(E_{-\infty})$ and $\mathbb{P}_{\pi_2}(E_0)$ in PSPACE. ◀

Proof of Theorem 17.2. Immediate from Corollary 12 and the polynomial-time decidability of distinguishability [7]. ◀

Towards a proof of Theorem 17.3, 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 [19].

Concerning $\mathbb{P}_{\pi_2}(E_{-\infty})$ (cf. Theorem 17.3), 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 20.** *There is a polynomial-time computable function that maps any instance of the mortality problem to an 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 an HMM (Q, Σ, Ψ) so that $\Phi(a)$ and $\Psi(a)$ 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 $\mathbb{P}_{\pi_2}(E_{-\infty}) = 0$. ◀

The proof that deciding whether $\mathbb{P}_{\pi_2}(E_0) = 1$ is PSPACE-hard is similarly based on mortality.

5 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 21.** *A* matrix system *is a triple $\mathcal{M} = (Q, \Sigma, \Psi)$ where Q is a finite set of states, Σ is a finite set of observations, and $\Psi : \Sigma \to \mathbb{R}_{\geq 0}^{Q \times Q}$ specifies the transitions. (Note that an HMM is a matrix system.) A* Lyapunov system *is a pair $\mathcal{S} = (\mathcal{M}, \rho)$ where $\mathcal{M} = (Q, \Sigma, \Psi)$ 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 $E = \{(q, r) \mid \sum_{a \in \Sigma} \Psi_{q,r}(a) > 0\}$ is strongly connected.*

We can identify the probability distribution ρ from this definition with the single-state HMM $(\{s\}, \Sigma, \Psi_\rho)$ where $\Psi_\rho(a)_{s,s} = \rho(a)$ for all $a \in \Sigma$. In this way, ρ produces a random infinite word from Σ^ω. The following lemma is known from [26].

▶ **Lemma 22** ([26])**.** *Let $((Q, \Sigma, \Psi), \rho)$ be a Lyapunov system. Then there is $\lambda \in \mathbb{R}$ such that, for all $q \in Q$, $\mathbb{P}_\rho$-a.s., either $e_q \Psi(w_n) = \vec{0}$ for some $n \in \mathbb{N}$ or the limit $\lim_{n\to\infty} \frac{1}{n} \ln \|e_q \Psi(w_n)\|$ exists and equals λ.*

For a Lyapunov system $\mathcal{S}$ we call $\lambda(\mathcal{S}) = \lambda$ from the lemma the *Lyapunov exponent* defined by $\mathcal{S}$. We prove the following theorem, which implies Theorem 8.

▶ **Theorem 23.** *Given an HMM (Q, Σ, Ψ) 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, the HMM (Q, Σ, Ψ) has at most $|Q|^2 + 1$ likelihood exponents.

In the rest of the section we provide more details on the construction underlying Theorem 23. As an intermediate concept (between the given HMM and the Lyapunov systems from Theorem 23) we define *generalized Lyapunov systems*.

First, for two matrix systems $\mathcal{M}_1 = (Q_1, \Sigma, \Psi_1)$ and $\mathcal{M}_2 = (Q_2, \Sigma, \Psi_2)$ with finite Q_1, Q_2, Σ and transitions $\Psi_1, \Psi_2 : \Sigma \to \mathbb{R}_{\geq 0}^{Q \times Q}$ we define the directed graph $G_{\mathcal{M}_1, \mathcal{M}_2} = (Q_1 \times Q_2, E)$ such that there is an edge from (q_1, q_2) to (r_1, r_2) if there is $a \in \Sigma$ with $\Psi_1(a)_{q_1,r_1} > 0$ and $\Psi_2(a)_{q_2,r_2} > 0$.

A *generalized Lyapunov system* is a triple $\mathcal{S} = (\mathcal{M}, \mathcal{H}, C)$ where $\mathcal{M} = (Q_1, \Sigma, \Psi_1)$ is a matrix system and $\mathcal{H} = (Q_2, \Sigma, \Psi_2)$ is a strongly connected HMM and $C \subseteq Q_1 \times Q_2$ is a bottom SCC of $G_{\mathcal{M},\mathcal{H}}$. Given a generalized Lyapunov system, one can efficiently compute an "equivalent" Lyapunov system:

▶ **Lemma 24.** *Let $\mathcal{S} = ((Q_1, \Sigma, \Psi_1), (Q_2, \Sigma, \Psi_2), C)$ 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 $\operatorname{supp}(\pi_1) \times \operatorname{supp}(\pi_2) \subseteq C$, we have $\mathbb{P}_{\pi_2}$-a.s. that either $\pi_1 \Psi_1(w_n) = \vec{0}$ for some $n \in \mathbb{N}$ or the limit $\lim_{n\to\infty} \frac{1}{n} \ln \|\pi_1 \Psi_1(w_n)\|$ exists and equals $\lambda(\mathcal{S})$.*
2. *One can compute in polynomial time a Lyapunov system $\mathcal{S}'$ such that $\lambda(\mathcal{S}) = \lambda(\mathcal{S}')$.*

Let $\mathcal{H} = (Q, \Sigma, \Psi)$ be an HMM. Let $R \subseteq Q \times Q$ be a (not necessarily bottom) SCC of the graph $G_{\mathcal{H},\mathcal{H}}$ 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 graph of $\sum_{a \in \Sigma} \Psi(a)$. We call such R a *right-bottom* SCC. Clearly there are at most

$|Q|^2$ right-bottom SCCs. Towards Theorem 23 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.

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.

Define $\Psi'(a,r)_{q,r} := \Psi(a)_{q,r}$ for all $a \in \Sigma$ and $q, r \in Q$. Then $(Q, \Sigma \times Q, \Psi')$ is an HMM, which is similar to $\mathcal{H}$, but which emits, in addition to an observation from Σ, also the next state. Since Q_R is a bottom SCC of the graph of $\sum_{a\in\Sigma} \Psi(a)$, the HMM $\mathcal{H}_2 := (Q_R, \Sigma \times Q_R, \Psi'_{|Q_R})$ is strongly connected. This HMM $\mathcal{H}_2$ will be used both in $\mathcal{S}^1_R$ and in $\mathcal{S}^2_R$.

Next, define $\overline{\Psi} : (\Sigma \times Q) \to [0,1]^{(Q\times Q)\times(Q\times Q)}$ by

$$\overline{\Psi}(a, r_2)_{(q_1,q_2),(r_1,r_2)} \;:=\; \Psi(a)_{q_1,r_1} \quad \text{for all } a \in \Sigma \text{ and } q_1, q_2, r_1, r_2 \in Q\,.$$

Now define $\mathcal{S}^1_R := (\mathcal{M}^1, \mathcal{H}_2, C^1)$, where $\mathcal{M}^1 := (R, \Sigma \times Q_R, \overline{\Psi}_{|R})$ 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 graph $G_{\mathcal{H},\mathcal{H}}$ that contains the "diagonal" vertices $(q,q) \in Q_R \times Q_R$, define $\mathcal{S}^2_R := (\mathcal{M}^2, \mathcal{H}_2, C^2)$, where $\mathcal{M}^2 := (R', \Sigma \times Q_R, \overline{\Psi}_{|R'})$ and $C^2 := \{((q_1,q_2),q_2) \mid (q_1,q_2) \in R'\}$.

For sets $U, V \subseteq Q \times Q$ let $U \longrightarrow_{G_{\mathcal{H},\mathcal{H}}} V$ denote that there are $u \in U$ and $v \in V$ such that v is reachable from u in $G_{\mathcal{H},\mathcal{H}}$. We are ready to state the following key technical lemma:

▶ **Lemma 25.** *Given an HMM (Q, Σ, Ψ), 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*

$$\{-\infty\} \cup \{\lambda(\mathcal{S}^1_R) - \lambda(\mathcal{S}^2_R) \mid R \in \mathcal{R},\ \mathrm{supp}(\pi_1) \times \mathrm{supp}(\pi_2) \longrightarrow_{G_{\mathcal{H},\mathcal{H}}} R\}\,.$$

Thus, $\Lambda_{\pi_1,\pi_2} \subseteq \{-\infty\} \cup \{\lambda(\mathcal{S}^1_R) - \lambda(\mathcal{S}^2_R) \mid R \in \mathcal{R},\ \mathrm{supp}(\pi_1) \times \mathrm{supp}(\pi_2) \longrightarrow_{G_{\mathcal{H},\mathcal{H}}} R\}$.

Proof sketch. Let π_1, π_2 be initial distributions. Very loosely speaking, we show in the appendix that on $\mathbb{P}_{\pi_2}$-almost every run w there is a right-bottom SCC R which "traps" "most" of the mass of $\pi_1\Psi(w_n)$ and $\pi_2\Psi(w_n)$. This can be made meaningful and formal using (the cross-product systems) $\mathcal{S}^1_R, \mathcal{S}^2_R$. We then show that on $\mathbb{P}_{\pi_2}$-almost every such run w, for both $i = 1, 2$, the limit $\lim_{n\to\infty} \frac{1}{n} \ln \|\pi_i\Psi(w_n)\|$ exists and equals $\lambda(\mathcal{S}^i_R)$ (or $\pi_1\Psi(w_n) = \vec{0}$ for some n). It follows that

$$\lim_{n\to\infty} \frac{1}{n} \ln L_n \;=\; \lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\pi_1\Psi(w_n)\|}{\|\pi_2\Psi(w_n)\|} \;=\; \lambda(\mathcal{S}^1_R) - \lambda(\mathcal{S}^2_R)\,. \qquad \blacktriangleleft$$

With Lemma 25 at hand, the proof of Theorem 23 is easy:

Proof of Theorem 23. 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 25 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 Lemma 24.2, for each $R \in \mathcal{R}$ one can compute in polynomial time an equivalent Lyapunov system. $\blacktriangleleft$

Theorem 23 allows us to represent the likelihood exponents of an HMM in terms of Lyapunov exponents. In general, approximating or even computing Lyapunov exponents is hard, but there are practical approximation algorithms using convex optimisation [27, 30].

6 Deterministic HMMs

In Sections 4 and 5 we have seen that the problems of representing/computing likelihood exponents and of computing their probabilities tend to be computationally difficult. In this section we study *deterministic* HMMs and show that this subclass leads to tractable problems. An HMM (Q, Σ, Ψ) is *deterministic* if, for all $a \in \Sigma$, all rows of $\Psi(a)$ contain at most one non-zero entry. Thus, for all $q \in Q$ and $w \in \Sigma^*$, we have $|\mathrm{supp}(e_q \Psi(w))| \leq 1$.

A useful observation is that the Markov chain $\mathcal{B} = (2^Q \times Q, T)$, which was defined before Lemma 19 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 26.** Consider the deterministic HMM (Q, Σ, Ψ) in Figure 3(a). Let $\pi_1 = e_{q_1}$

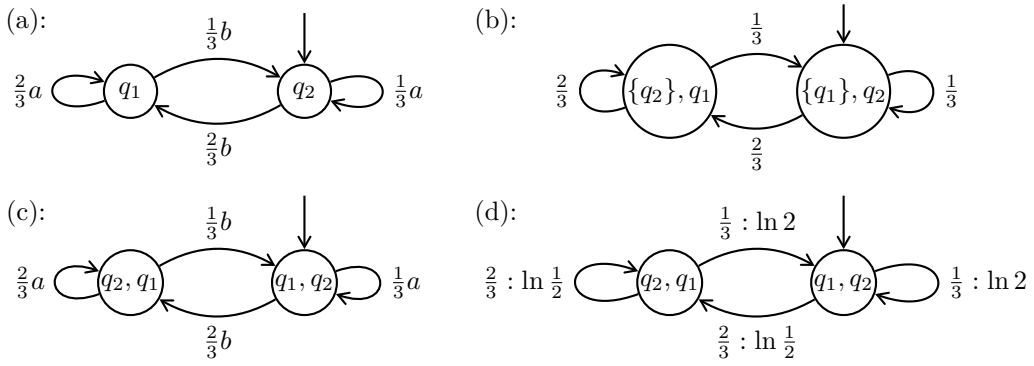


Figure 3 Cross-product constructions for a deterministic HMM.

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 $\mathcal{B}$ is shown in Figure 3(b). Let us add back the observations that gave rise to the transitions in $\mathcal{B}$, and for simplicity drop the set brackets in the left component of states. We obtain the HMM in Figure 3(c). With this HMM we may keep track of the exact likelihood ratio. For example, suppose that the word aba is emitted, so that $L_3 = \frac{\|e_{q_1}\Psi(aba)\|}{\|e_{q_2}\Psi(aba)\|} = \frac{1}{2}$ and $\mathrm{supp}(e_{q_1}\Psi(aba)) = \{q_2\}$ and $\mathrm{supp}(e_{q_2}\Psi(aba)) = \{q_1\}$. 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 $\frac{\Psi_{q_2,q_1}(b)}{\Psi_{q_1,q_2}(b)} = 2$, 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 3(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)

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

In general there may again be several likelihood exponents, including $-\infty$ and 0. For the rest of the section, let $\mathcal{H} = (Q, \Sigma, \Psi)$ be a deterministic HMM. Motivated by Example 26, define an HMM $\mathcal{A} = ((Q \times Q) \cup s_\perp, \hat{\Sigma}, \hat{\Psi})$, where $s_\perp$ is a fresh state, and

$$\hat{\Sigma} := \left\{ \ln \frac{\Psi(a)_{q_1,r_1}}{\Psi(a)_{q_2,r_2}} \in [-\infty, \infty) \;\middle|\; a \in \Sigma,\ q_1, r_1, q_2, r_2 \in Q,\ \Psi(a)_{q_2,r_2} \neq 0 \right\} \cup \{-\infty\}$$

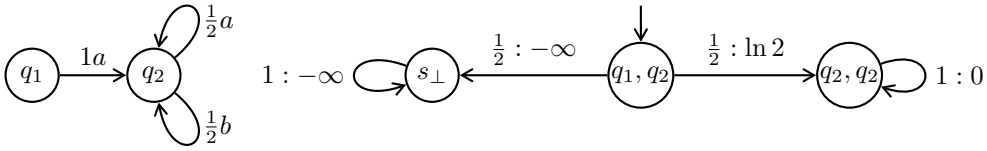
$$\hat{\Psi}(\hat{a})_{(q_1,q_2),(r_1,r_2)} := \sum \left\{ \Psi(a)_{q_2,r_2} \;\middle|\; a \in \Sigma : \hat{a} = \ln \frac{\Psi(a)_{q_1,r_1}}{\Psi(a)_{q_2,r_2}} \right\} \quad \text{for } \hat{a} \neq -\infty$$

$$\hat{\Psi}(-\infty)_{(q_1,q_2),s_\perp} := \sum \left\{ \Psi(a)_{q_2,r_2} \;\middle|\; a \in \Sigma,\ r_2 \in Q : \textstyle\sum_{r_1 \in Q} \Psi(a)_{q_1,r_1} = 0 \right\}$$

$$\hat{\Psi}(-\infty)_{s_\perp,s_\perp} := 1\,.$$

Note that the embedded Markov chain of $\mathcal{A}$ is similar to the Markov chain $\mathcal{B}$ from Lemma 19: states $(\{q_1\}, q_2)$ in $\mathcal{B}$ are called (q_1, q_2) in $\mathcal{A}$, the states $(\emptyset, q)$ in $\mathcal{B}$ are subsumed by the state $s_\perp$ of $\mathcal{A}$, and the states (S, q) in $\mathcal{B}$ with $|S| > 1$ are not represented in $\mathcal{A}$. The observations in $\hat{\Sigma} \subseteq [-\infty, \infty)$ track the log-likelihood ratio.

▶ **Example 27.** Consider the HMM $\mathcal{H}$ 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 $\nu_{(r_1,r_2)} := \sum_{\hat{a} \in \hat{\Sigma}} \|e_{(r_1,r_2)} \hat{\Psi}(\hat{a})\| \cdot \hat{a}$. 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 28.** *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 $\mathbb{P}_{\pi_2}(E_\ell) = \mathbb{P}_\iota(\{\text{visit bottom SCC } C \text{ with } \ell(C) = \ell\})$.*

The proof is essentially the same as in Lemma 19.3. This gives us the following result.

▶ **Theorem 29.** *Given a deterministic HMM (Q, Σ, Ψ) 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 sketch. The theorem follows mostly from Lemma 28, with the slight complication that for part 2 we have to check numbers of the form $\sum_i x_i \ln y_i$ (where $x_i, y_i \in \mathbb{Q}$) for equality. But this can be done in polynomial time as shown in [15]. ◀

7 Conclusions

We have shown that the performance 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 [18, 5, 16, 17]. Efficient approximation of likelihood exponents, in theory or praxis, remains an open problem.

References

1 P. Ailliot, C. Thompson, and P. Thomson. Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. *Journal of the Royal Statistical Society*, 58(3):405–426, 2009.

2 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)*, volume 150 of *LIPIcs*, pages 29:1–29:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.FSTTCS.2019.29.

3 M. Alexandersson, S. Cawley, and L. Pachter. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, 13:469–502, 2003.

4 N. Bertrand, S. Haddad, and E. Lefaucheux. Accurate approximate diagnosability of stochastic systems. In *Proceedings of Language and Automata Theory and Applications (LATA)*, volume 9618 of *Lecture Notes in Computer Science*, pages 549–561. Springer, 2016. doi:10.1007/978-3-319-30000-9_42.

5 B. Chen and P. Willett. Detection of hidden Markov model transient signals. *IEEE Transactions on Aerospace and Electronic Systems*, 36(4):1253–1268, 2000. doi:10.1109/7.892673.

6 F.-S. Chen, C.-M. Fu, and C.-L. Huang. Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing*, 21(8):745–758, 2003.

7 T. Chen and S. 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)*, pages 33:1–33:10, Vienna, Austria, 2014.

8 G.A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1):79–94, 1989.

9 C. Cortes, M. Mohri, and A. Rastogi. L_p distance and equivalence of probabilistic automata. *International Journal of Foundations of Computer Science*, 18(04):761–779, 2007.

10 M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, 46(4):886–902, April 1998.

11 O. Darwin and S. Kiefer. On the sequential probability ratio test in hidden Markov models, 2022. arXiv:2207.14088.

12 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)*, pages 592–600. Springer, 2017.

13 R. Durbin. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

14 S.R. Eddy. What is a hidden Markov model? *Nature Biotechnology*, 22(10):1315–1316, October 2004.

15 K. Etessami, A. Stewart, and M. Yannakakis. A note on the complexity of comparing succinctly represented integers, with an application to maximum probability parsing. *ACM Trans. Comput. Theory*, 6(2):9:1–9:23, 2014. doi:10.1145/2601327.

16 C.-D. Fuh. SPRT and CUSUM in hidden Markov models. *The Annals of Statistics*, 31(3):942–977, 2003. doi:10.1214/aos/1056562468.

17 E. Grossi and M. Lops. Sequential detection of Markov targets with trajectory estimation. *IEEE Transactions on Information Theory*, 54(9):4144–4154, 2008. doi:10.1109/TIT.2008.928261.

18 B.-H. Juang and L. R. Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64(2):391–408, 1985. doi:10.1002/j.1538-7305.1985.tb00439.x.

19 J.-Y. Kao, N. Rampersad, and J. Shallit. On NFAs where all states are final, initial, or both. *Theoretical Computer Science*, 410(47):5010–5021, 2009. doi:10.1016/j.tcs.2009.07.049.

20 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)*, volume 6806 of *LNCS*, pages 526–540. Springer, 2011.

21 S. Kiefer and A.P. Sistla. Distinguishing hidden Markov chains. In *Proceedings of the 31st Annual Symposium on Logic in Computer Science (LICS)*, pages 66–75, New York, USA, 2016. ACM.

22 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. *Journal of Molecular Biology*, 305(3):567–580, 2001.

23 M. Kwiatkowska, G. Norman, and D. Parker. PRISM 4.0: Verification of probabilistic real-time systems. In *Proceedings of Computer Aided Verification (CAV)*, volume 6806 of *LNCS*, pages 585–591. Springer, 2011.

24 R. Langrock, B. Swihart, B. Caffo, N. Punjabi, and C. Crainiceanu. Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in medicine*, 32, August 2013. doi:10.1002/sim.5747.

25 A. Paz. *Introduction to Probabilistic Automata (Computer Science and Applied Mathematics)*. Academic Press, Inc., Orlando, FL, USA, 1971.

26 V.Yu. Protasov. Asymptotics of products of nonnegative random matrices. *Functional Analysis and Its Applications*, 47:138–147, 2013.

27 V.Yu. Protasov and R.M. Jungers. Lower and upper bounds for the largest Lyapunov exponent of matrices. *Linear Algebra and its Applications*, 438(11):4448–4468, 2013. doi:10.1016/j.laa.2013.01.027.

28 L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

29 M.P. Schützenberger. On the definition of a family of automata. *Information and Control*, 4(2):245–270, 1961.

30 D. Sutter, O. Fawzi, and R. Renner. Bounds on Lyapunov exponents via entropy accumulation. *IEEE Transactions on Information Theory*, 67(1):10–24, 2021. doi:10.1109/TIT.2020.3026959.

31 W.-G. Tzeng. A polynomial-time algorithm for the equivalence of probabilistic automata. *SIAM J. Comput.*, 21(2):216–227, April 1992.

32 A. Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945. doi:10.1214/aoms/1177731118.

33 A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948. URL: http://www.jstor.org/stable/2235638.