Distinguishing Hidden Markov Chains*

Stefan Kiefer
University of Oxford, UK

A. Prasad Sistla
University of Illinois at Chicago, USA

ABSTRACT

Hidden Markov Chains (HMCs) are commonly used mathematical models of probabilistic systems. They are employed in various fields such as speech recognition, signal processing, and biological sequence analysis. Motivated by applications in stochastic runtime verification, we consider the problem of distinguishing two given HMCs based on a single observation sequence that one of the HMCs generates. More precisely, given two HMCs and an observation sequence, a distinguishing algorithm is expected to identify the HMC that generates the observation sequence. Two HMCs are called distinguishable if for every $\varepsilon > 0$ there is a distinguishing algorithm whose error probability is less than ε. We show that one can decide in polynomial time whether two HMCs are distinguishable. Further, we present and analyze two distinguishing algorithms for distinguishable HMCs. The first algorithm makes a decision after processing a fixed number of observations, and it exhibits two-sided error. The second algorithm processes an unbounded number of observations, but the algorithm has only one-sided error. The error probability, for both algorithms, decays exponentially with the number of processed observations. We also provide an algorithm for distinguishing multiple HMCs.

CCS Concepts

•**Theory of computation** → **Random walks and Markov chains;** *Probabilistic computation;* •**Mathematics of computing** → *Computing most probable explanation;*

Keywords

Hidden Markov chains; Labelled Markov chains; monitors

1. INTRODUCTION

Hidden Markov Chains (HMCs) are commonly used mathematical models of probabilistic systems. They are specified by a Markov Chain, capturing the probabilistic behavior of a system, and an observation function specifying the outputs generated from each of its states. Figure 1 depicts two example HMCs H_1, H_2, with observations a and b. We consider finite-state HMCs in this paper. An HMC randomly generates a (conceptually infinite) string of observations. The states producing the observations are not observable (note that s_0 and s_1 output the same observation a in the example). This motivates the term *hidden*.

HMCs are widely employed in fields such as speech recognition (see [22] for a tutorial), gesture recognition [6], musical score following [23], signal processing [9], and climate modeling [1]. HMCs are heavily used in computational biology [12], more specifically in DNA modeling [8] and biological sequence analysis [11], including protein structure prediction [18], detecting similarities in genomes [14] and gene finding [2]. Following [19], applications of HMCs are based on two basic problems, cf. [13, Chapter 2]: The first one is, given an observation string and an HMC, what is the most likely sequence of states that produced the string? This is useful for areas like speech recognition, see [22] for efficient algorithms based on dynamic programming. The second problem is, given an observation string and multiple HMCs, identify the HMC that is most likely to produce the observation. This is used for classification.

The second problem raises a fundamental question, which we address in this work: Given two HMCs, and assuming that one of them produces a random single observation sequence, is it even possible to identify the producing HMC with a high probability? And if yes, how many observations in that observation sequence are needed? At its heart, this question is about comparing two HMCs in terms of their (distributions on) observation sequences. To make this more precise, let a *monitor* for two given HMCs H_1, H_2 be an algorithm that reads (increasing prefixes of) a single observation sequence, and at some point outputs "H_1" or "H_2". The *distinguishability* problem asks for two given HMCs H_1, H_2, whether for all $\varepsilon > 0$ there is a monitor such that for both $i = 1, 2$, if the monitor reads a random observation sequence produced by H_i, then with probability at least $1 - \varepsilon$ the monitor outputs "H_i".

A related problem is *equivalence* of HMCs. Two HMCs are called *equivalent* if they produce the same (prefixes of) observation sequences with the same probability. Equivalence of HMCs has been well-studied and can be decided in polynomial time, using algorithms based on linear algebra,

*This is the full version of a LICS'16 paper.

LICS '16, July 05 - 08, 2016, New York, NY, USA

ISBN 978-1-4503-4391-6/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2933575.2933608

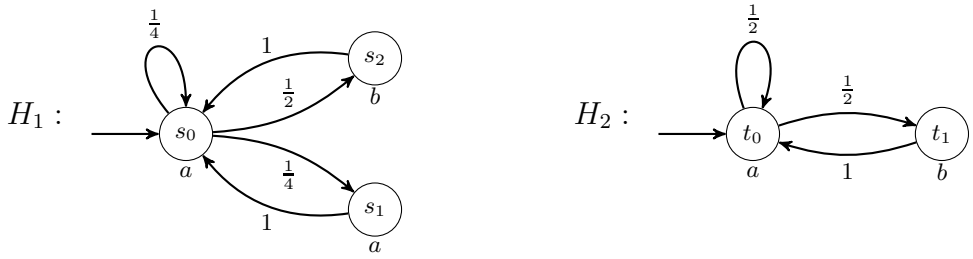

Figure 1: Two HMCs. Here H_1 and H_2 are distinguishable (see Example 2) and hence not equivalent.

see e.g. [15, 29, 10]. The exact relation between equivalence and distinguishability depends on whether a monitor has access to a single random observation sequence or to multiple such sequences.

(1) Consider first a notion of a "monitor" that has access to *several* random observation sequences, each generated starting from the same initial state. Call this a *multi-monitor*. If the two given HMCs are equivalent then even a multi-monitor can only guess. Now assume the two HMCs are not equivalent. It is known (see e.g. [29]) that then there exists a linear-length prefix of the observation sequence that is more likely in one HMC than in the other HMC. A multi-monitor could exploit the law of large numbers and only count how often that particular observation prefix occurs. Hence for multi-monitors, distinguishability and non-equivalence coincide.

(2) Consider now a monitor that has access to only a *single* random observation sequence. Here, non-equivalence does not imply distinguishability: loosely speaking, for some HMCs it is the case that while the observation prefix is increasing, the evidence added by each new observation does not help the monitor enough to make up its mind about which HMC produces the sequence. Figure 2 shows an example of two HMCs that are neither equivalent nor distinguishable. (On the other hand, the HMCs in Figure 1 are not equivalent, but are distinguishable as shown later in Section 3).

We assume in the rest of the paper that a monitor has access to only a single random observation sequence. This is the more natural version of the problem, both from the point of view of the motivation mentioned above and from our application in stochastic runtime monitoring.

We prove that the distinguishability problem is decidable in polynomial time. We establish this result by showing that two HMCs are distinguishable if and only if their *total variation distance* is equal to 1. This distance measure for HMCs was studied in [7], and a polynomial-time algorithm for deciding whether the distance of two HMCs is 1 was given there. That polynomial-time algorithm includes a mechanism for checking whether two given HMCs are equivalent (but also needs other ingredients).

It is important to note that deciding distinguishability does not readily provide a family of monitors as required by the definition of distinguishability; it only guarantees their existence. Developing a family of monitors (one for any desired error bound $\varepsilon > 0$) requires more insights. Inspired by the area of *sequential analysis* [30], we design monitors that track the *likelihood ratio* of the sequence of observations. However, estimating the error probability of the monitors is challenging, since one needs a bound on the change of the likelihood ratio per observation. Unfortunately, such a bound does not exist for HMCs in general, not even on the difference of the *log*-likelihood ratio (see Example 6). Hence, in this paper we take a different route: We consider a different class of monitors that translate the given random observation sequence into a certain kind of non-homogenous "random walk" with *bounded* step size. This allows us to employ martingale techniques, specifically Azuma's inequality, to prove error bounds that decay exponentially with the number of observations the monitor makes. Then we show that the error bounds from a random-walk monitor carry over to a likelihood-based monitor.

More specifically, we present two likelihood-based monitors for distinguishable HMCs. The first one makes a decision after reading a fixed number of observation symbols. This number is chosen depending on the desired error bound: we show that for an error probability ε it suffices to read the prefix of length $C \log \frac{1}{\varepsilon}$, where $C > 0$ is a polynomial-time computable constant. This error is two-sided, i.e., the monitor may mistake H_1 for H_2 and vice versa.

The second monitor has only one-sided error: observation sequences from H_1 are almost always (i.e., with probability 1) recognized as stemming from H_1. However, on sequences generated by H_2, with high probability the monitor never gives an answer. This is useful in applications such as runtime verification (see Section 6). The expected number of observations from H_1 that the monitor processes before giving its decision is $O(\log \frac{1}{\varepsilon})$, while ensuring an error probability of at most ε on observations from H_2. For this class of monitors, we have a polynomial-time algorithm that computes an $O(\log \frac{1}{\varepsilon})$ upper bound on the expected number of observations from H_1 before a decision is given.

Main Contributions.

- We show that the distinguishability problem can be decided in polynomial time (Section 3).
- We design two classes of likelihood-based monitors that accomplish the following tasks ($\varepsilon > 0$ is an error bound):
  (1) After $O(\log \frac{1}{\varepsilon})$ observations (the exact number can be efficiently computed from the given HMCs) the first monitor class provides a guess about the source of the observations, such that the probability that the guess is wrong is at most ε (Sec-

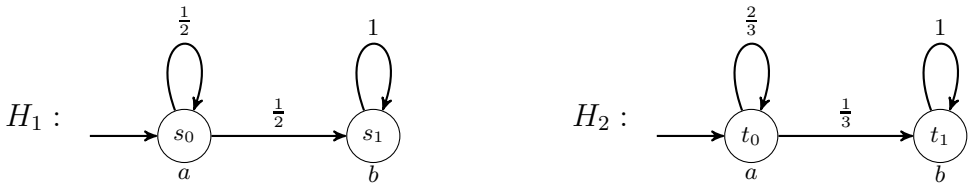

Figure 2: Two HMCs. Here H_1 and H_2 are not distinguishable but not equivalent.

tion 4.2). This can be extended to more than two HMCs (Section 4.4).

(2) For the second monitor class, if H_1 produces the observation sequence then the monitor raises an alarm almost surely, and after an *expected* number of $O(\log \frac{1}{\varepsilon})$ observations (such an upper bound can be efficiently computed from the given HMCs and ε); if H_2 produces the observation sequence then, with probability at least $1-\varepsilon$, the monitor never raises an alarm (Section 4.3).

- We apply our results to stochastic runtime verification, where a monitor should distinguish correct and faulty behaviour of a *single* stochastic system. This yields polynomial-time decidability of monitorability as defined in [26], as well as efficient runtime monitors for stochastic systems, see Section 6.

Missing proofs can be found in the appendix.

Related Work. The area of *sequential analysis* in statistics, pioneered by Wald (see [30]), deals with the problem of hypothesis testing using repeated and unbounded sampling. A line of work going back to Phatarfod [21, 28, 24] investigated the application of sequential analysis, more specifically the sequential probability ratio test, to Markov chains. Similar to our work, the goal in the above works is to identify a Markov chain among several, in this case using likelihood ratios. A monitor algorithm is derived by keeping track of likelihood ratios: it gives notice once the likelihood ratio drops below or exceeds some fixed threshold. One problem with this approach is that error probabilities can only be estimated—not bounded—by the heuristic assumption that the excess over the threshold is not big. This assumption is not always true. A more important difference from our work is that the observation in each state equals the state, in other words, the Markov chains are not hidden.

There is early related work that is more specific to HMCs. The paper [16] aims at *measuring* a certain distance between two HMCs by running one of them. This is in spirit close to our work, as a positive distance in their sense could be transformed to a monitor. However, the authors place strong assumptions on the Markov chains, in particular ergodicity. If this assumption is removed, their distance can be different for different runs, and the existence of a lower bound on the possible distances is unclear.

Work by Alur et al. [3] also aims at distinguishing probabilistic models, but there are important differences. First, they consider Markov Decision Processes rather than Markov chains, i.e., they consider *strategies* to distinguish two such processes, which is a more general, and computationally harder problem (they show PSPACE- and EXPTIME-completeness results). Second, their problems are defined such that the exact values of the transition probabilities is unimportant. In our case this is different.

The work in [19] deals with comparing two HMCs in terms of various distance measures. Among other results, they show NP-hardness of computing and approximating the ℓ_1-distance. The HMCs considered there generate distributions on *finite* strings of observations, as each HMC has a dedicated end state, reached with probability 1, where the HMC "stops". Such HMCs form a subclass of HMCs, whereas we consider general HMCs.

Our work on distinguishability is inspired by the work on monitorability that was defined in [26]. In [26, Section 4.1] a notion of *strong monitorability* is proposed and it is shown that deciding it is PSPACE-complete. By our results in Section 6, strong monitorability corresponds to a stronger form of distinguishability, so the latter is PSPACE-complete as well. In light of this it might be surprising that (general) distinguishability turns out to be decidable in polynomial time. In [26] it was wrongly claimed that *monitorability* is undecidable for finite-state systems. Our result not only shows that it is decidable, but also gives a polynomial-time decision procedure.

Our work on exponentially decaying monitors is inspired by the exponentially converging monitorable systems defined in [27]. The algorithms presented there are for a very restricted class of HMCs, whereas our monitors work for all pairs of distinguishable HMCs.

Closely related to some of our results is a very recent work by Bertrand et. al. [5]. This paper also exploits the results of [7] to obtain polynomial-time decidability of "AA-diagnosability" of stochastic systems, a problem related to monitorability (Section 6). Although the technical report of our work had been available [17], the results in [5] were obtained independently and are largely orthogonal to ours: whereas we focus on constructing specific monitors with computable error bounds, they investigate the decidability and complexity of several variants of diagnosability.

2. DEFINITIONS

Notation. For a countable set S, a probability distribution ψ over S is a function $\psi : S \to [0,1]$ such that $\sum_{s\in S} \psi(s) = 1$. For an element $s \in S$, we let δ_s denote the unique distribution with $\delta_s(s) = 1$. We let $Distr(S)$ denote the set of all distributions over S. We let S^*, S^ω respectively denote the set of finite sequences (*strings*) and the set of infinite sequences of symbols from S. If S is a finite set then we let $|S|$ denote its cardinality. For any $u \in S^*$, we let $|u|$ denote its length. For any real number x, we let $|x|$ denote its absolute value.

Hidden Markov Chains. A Markov chain is a triple $G = (S, R, \phi)$ where S is a set of states, $R \subseteq S \times S$, and

$\phi : R \to (0,1]$ is such that $\sum_{t:(s,t)\in R} \phi(s,t) = 1$ for all $s \in S$. A Markov chain G and an initial state $s \in S$ induce a probability measure, denoted by $\mathcal{P}_s$, on measurable subsets of $\{s\}S^\omega$ in the usual way: more precisely, we consider the σ-algebra generated by the cylinder sets $\{s_0 s_1 \cdots s_n\}S^\omega$ for $n \geq 0$ and $s_0 = s$ and $s_i \in S$, with the probability measure $\mathcal{P}_s$ such that

$$\mathcal{P}_s(\{s_0 s_1 \cdots s_n\}S^\omega) = \prod_{i=1}^{n} \phi(s_{i-1}, s_i)\,.$$

Let Σ be a finite set. A *Hidden Markov Chain (HMC)*, with observation alphabet Σ, is a triple (G, O, s_0), where $G = (S, R, \phi)$ is a Markov chain, and $O : S \to \Sigma$ is the observation function, and $s_0 \in S$ is the initial state. We may write $\mathcal{P}$ for $\mathcal{P}_{s_0}$. For $\mathcal{L} \subseteq \Sigma^\omega$ we define the inverse observation function

$$[\mathcal{L}] := \{s_0 s_1 \cdots \in S^\omega \mid O(s_0)O(s_1)\cdots \in \mathcal{L}\}\,.$$

Monitors. A *monitor* $M : \Sigma^* \to \{\bot, 1\}$ is a computable function with the property that, for any $u \in \Sigma^*$, if $M(u) = 1$ then $M(uv) = 1$ for every $v \in \Sigma^*$. Let $\mathcal{L}(M) \subseteq \Sigma^\omega$ denote the set of infinite sequences that have a prefix u with $M(u) = 1$. (Intuitively, $\mathcal{L}(M)$ is the set of observation sequences which the monitor decides to have been generated by the first HMC among a pair of such HMCs.) Given an HMC, the event $[\mathcal{L}(M)]$ is measurable, as it is a countable union of cylinder sets.

Distinguishability. Given two HMCs H_1, H_2 with the same observation alphabet Σ, we write $\mathcal{P}_1, \mathcal{P}_2, [\cdot]_1, [\cdot]_2$ for their associated probability measures and inverse observation functions. HMCs H_1, H_2 are called *distinguishable* if for every $\varepsilon > 0$ there exists a monitor M such that

$$\mathcal{P}_1([\mathcal{L}(M)]_1) \;\geq\; 1-\varepsilon \quad \text{and} \quad \mathcal{P}_2([\mathcal{L}(M)]_2) \;\leq\; \varepsilon\,.$$

3. POLYNOMIAL-TIME DECIDABILITY OF THE DISTINGUISHABILITY PROBLEM

For two HMCs H_1, H_2 define the *(total variation) distance* between H_1 and H_2, denoted by $d(H_1, H_2)$, as follows:

$$d(H_1, H_2) := \sup_{E \subseteq \Sigma^\omega} |\mathcal{P}_1([E]_1) - \mathcal{P}_2([E]_2)|\ ,$$

where the supremum ranges over all *measurable* subsets of Σ^ω. It is shown in [7] that the supremum is in fact a maximum. In particular, if $d(H_1, H_2) = 1$ then there exists a measurable set $E \subseteq \Sigma^\omega$ with $\mathcal{P}_1([E]_1) = 1$ and $\mathcal{P}_2([E]_2) = 0$. We show:

PROPOSITION 1. *HMCs H_1, H_2 are distinguishable if and only if $d(H_1, H_2) = 1$.*

PROOF. Let H_1, H_2 be two given HMCs. We show that H_1, H_2 are distinguishable if and only if $d(H_1, H_2) = 1$.

- *"if"*: Let $d(H_1, H_2) = 1$. Choose $\varepsilon > 0$ arbitrarily. It follows from [7, Theorem 7] and the discussion after [7, Proposition 5] that there are $k \in \mathbb{N}$ and $W \subseteq \Sigma^k$ such that

  $$\mathcal{P}_1([W\Sigma^\omega]_1) \geq 1-\varepsilon \quad \text{and} \quad \mathcal{P}_2([W\Sigma^\omega]_2) \leq \varepsilon\,.$$

  Construct a monitor M that outputs 1 after having read a string in W. Then we have $\mathcal{L}(M) = W\Sigma^\omega$. It follows that

  $$\mathcal{P}_1([\mathcal{L}(M)]_1) \geq 1-\varepsilon \quad \text{and} \quad \mathcal{P}_2([\mathcal{L}(M)]_2) \leq \varepsilon\,.$$

  Since ε was chosen arbitrarily, the HMCs H_1, H_2 are distinguishable.

- *"only if"*: Let H_1, H_2 be distinguishable, i.e., for every $\varepsilon > 0$ there exists a monitor M_ε such that

  $$\mathcal{P}_1([\mathcal{L}(M_\varepsilon)]_1) \geq 1-\varepsilon \quad \text{and} \quad \mathcal{P}_2([\mathcal{L}(M_\varepsilon)]_2) \leq \varepsilon\,.$$

  Then we have:

  $$\begin{aligned} d(H_1, H_2) &= \sup_{E \subseteq \Sigma^\omega} |\mathcal{P}([E]_1) - \mathcal{P}([E]_2)| \\ &\geq \sup_{\varepsilon > 0} \big(\mathcal{P}([\mathcal{L}(M_\varepsilon)]_1) - \mathcal{P}([\mathcal{L}(M_\varepsilon)]_2)\big) \\ &\geq \sup_{\varepsilon > 0} (1 - 2\varepsilon) \;=\; 1 \end{aligned}$$

This concludes the proof. □

It follows that HMCs H_1, H_2 are distinguishable iff there is a *distinguishing event*, i.e., a set $E \subseteq \Sigma^\omega$ with $\mathcal{P}_1([E]_1) = 1$ and $\mathcal{P}_2([E]_2) = 0$.

EXAMPLE 2. *Consider the HMCs H_1, H_2 from Figure 1. By computing the stationary distributions, one can show that, the distinguishing event E is given by*

$$E = \{\sigma_1\sigma_2\cdots \in \Sigma^\omega \mid \lim_{n\to\infty} \frac{f(n)}{n} = 5/7\}\ ,$$

where $f(n)$ denotes the number of occurrences of a in the prefix $\sigma_1\sigma_2\cdots\sigma_n$, is a distinguishing event for H_1, H_2. Hence H_1, H_2 are distinguishable. Here, counting the frequencies of the observations symbols suffices for distinguishing two distinguishable HMCs. In general, this is not true: the order *of observations may matter.* □

Proposition 1 implies the following theorem:

THEOREM 3. *One can decide in polynomial time whether given HMCs H_1, H_2 are distinguishable.*

PROOF. In [7, Algorithm 1 and Theorem 21] it is shown that, given two HMCs H_1, H_2, one can decide in polynomial time whether $d(H_1, H_2) = 1$. (The algorithm given there solves n_1 linear programs, each with $n_1 + n_2$ variables, where n_1, n_2 is the number of states in H_1, H_2, respectively.) Then the result follows from Proposition 1. □

Distinguishing events cannot in general be defined by monitors, as a monitor can reject an observation sequence only on the basis of a finite prefix. Moreover, the decision algorithm for Theorem 3 can assure the *existence* of a monitor for two given HMCs, but the decision algorithm does not provide useful monitors. That is the subject of the next section.

4. MONITORS

In this section, we present concrete monitors, with error bounds. To this end we give some additional definitions in Section 4.1, where we also explain how monitors can keep track of certain conditional distributions. We also introduce "profiles", a key concept for our proofs of error bounds. In Sections 4.2 and 4.3 we present monitors for distinguishable HMCs with two-sided and one-sided error, respectively. In

Section 4.4 we provide a monitor for distinguishing among multiple HMCs.

For $i = 1, 2$, let $H_i = (G_i, O_i, s_{i,0})$ be two HMCs with the same observation alphabet Σ, where $G_i = (S_i, R_i, \phi_i)$. Without loss of generality we assume $S_1 \cap S_2 = \emptyset$. Let $m := |S_1| + |S_2|$. We fix H_1, H_2 and m throughout the section.

4.1 Keeping Track of Probabilities and Profiles

Let $i \in \{1, 2\}$ and $\psi \in \mathit{Distr}(S_i)$. For $u \in \Sigma^*$ define

$$pr_i(\psi, u) := \sum_{s \in S_i} \psi(s) \cdot \mathcal{P}_{i,s}([u\Sigma^\omega]_i) \,.$$

Intuitively, $pr_i(\psi, u)$ is the probability that the string u is output by HMC H_i starting from the initial distribution ψ. For $W \subseteq \Sigma^m$ we also define

$$pr_i(\psi, W) := \sum_{u \in W} pr_i(\psi, u) \,,$$

which is the probability that H_i outputs a string in W starting from distribution ψ. For $u \in \Sigma\Sigma^*$ and $s, t \in S_i$ define

$$sub_i(s, u, t) := \mathcal{P}_{i,s}([u\Sigma^\omega]_i \cap S_i^{|u|-1}\{t\}S_i^\omega) \,.$$

Intuitively, $sub_i(s, u, t)$ is the probability that H_i outputs u and is then in state t, starting from state s. We have:

$$pr_i(\psi, u) = \sum_{s \in S_i} \psi(s) \cdot \sum_{t \in S_i} sub_i(s, u, t) \tag{1}$$

For any $s, r \in S_i$ and $u \in \Sigma\Sigma^*$ and $a \in \Sigma$ we have:

$$sub_i(s, ua, r) = \begin{cases} \sum_{t \in S_i} sub_i(s, u, t)\phi_i(t, r) & \text{if } O_i(r) = a \\ 0 & \text{otherwise} \end{cases}$$

So if a monitor has kept track of the values $sub_i(s, u, t)_{s,t \in S_i}$ for a prefix u of an observation sequence, it can, upon reading the next observation a, efficiently compute $sub_i(s, ua, t)_{s,t \in S_i}$ and, by (1), also $pr_i(\delta_{s_{i,0}}, ua)$.

For $u \in \Sigma^*$ define the *likelihood ratio*

$$lr(u) := \frac{pr_2(\delta_{s_{2,0}}, u)}{pr_1(\delta_{s_{1,0}}, u)} \,.$$

Finally, for $u \in \Sigma\Sigma^*$ with $pr_i(\psi, u) > 0$, define the distribution $cd_i(\psi, u)$ (which stands for "conditional distribution") as follows:

$$cd_i(\psi, u)(t) := \frac{1}{pr_i(\psi, u)} \cdot \sum_{s \in S_i} \psi(s) \cdot sub_i(s, u, t) \quad \text{for } t \in S_i$$

Intuitively, $cd_i(\psi, u)(t)$ is the conditional probability that H_i is in state t given that it has output u and started from ψ. As explained above, a monitor can efficiently keep track of $lr(u)$ and $cd_i(\psi, u)$.

We say that a pair of distributions $(\psi_1, \psi_2) \in \mathit{Distr}(S_1) \times \mathit{Distr}(S_2)$ is *reachable* in (H_1, H_2) if there is $u \in \Sigma\Sigma^*$ with $\psi_i = cd_i(\delta_{s_{i,0}}, u)$ for $i = 1, 2$. A *profile* for H_1, H_2 is a pair $(\mathcal{A}, c)$ such that $\mathcal{A} : \mathit{Distr}(S_1) \times \mathit{Distr}(S_2) \to 2^{\Sigma^m}$ and $c \in (0, 1]$ and

$$pr_1(\psi_1, \mathcal{A}(\psi_1, \psi_2)) - pr_2(\psi_2, \mathcal{A}(\psi_1, \psi_2)) \geq c$$

holds for all reachable pairs (ψ_1, ψ_2) of distributions. For the monitors presented in this section the following proposition is crucial.

PROPOSITION 4. *Let HMCs H_1, H_2 be distinguishable. Then there is a number $c > 0$, computable in time polynomial in the sizes of H_1, H_2, such that there is a profile $(\mathcal{A}, c)$.*

4.2 Monitors with Two-Sided Error

In this and the next subsection, we assume that H_1, H_2 are distinguishable, and fix a profile $(\mathcal{A}, c)$. The monitors of this subsection take an observation sequence as input, and at some point output a value from $\{1, 2, 3\}$ indicating a decision regarding which of the two HMCs generated the observations. An output of 3 indicates that neither of the HMCs could have generated it. The monitors of this subsection have two-sided errors: the answers 1 or 2 may be wrong (with a small probability).

We define a likelihood-based monitor M_2 (the subscript denotes two-sided error) as follows. Monitor M_2 runs in *phases*; in each phase, the monitor receives m observations. The monitor runs at most N phases, where $N \in \mathbb{N}$ is a parameter fixed in advance: choosing a larger N leads to smaller error probabilities. After reading an observation sequence u of length $N \cdot m$, it computes the likelihood ratio $lr(u)$. Monitor M_2 outputs 1 if $lr(u) < 1$, and 2 if $lr(u) > 1$. It may output either 1 or 2 if $lr(u) = 1$. Monitor M_2 needs no access to the function $\mathcal{A}$.

The following theorem says that the observation sequences for which monitor M_2 outputs 1 are much more likely to be generated by H_1. By symmetry, the observation sequences for which M_2 outputs 2 are much more likely to be generated by H_2.

THEOREM 5. *Consider the monitor M_2 that reads the first $N \cdot m$ observations. Let $\mathcal{L}(M_2) \subseteq \Sigma^\omega$ be the set of observation sequences for which M_2 outputs 1. Then we have*

$$\mathcal{P}_1([\mathcal{L}(M_2)]_1) - \mathcal{P}_2([\mathcal{L}(M_2)]_2) \geq 1 - 2\exp\left(-\frac{c^2}{18} \cdot N\right) \,.$$

Hence,

$$\mathcal{P}_1([\mathcal{L}(M_2)]_1) \;\geq\; 1 - 2\exp\left(-\frac{c^2}{18} \cdot N\right) \quad \textit{and}$$

$$\mathcal{P}_2([\mathcal{L}(M_2)]_2) \;\leq\; 2\exp\left(-\frac{c^2}{18} \cdot N\right) \,.$$

Proving the bounds of Theorem 5 is challenging due to the following reasons. For $k \geq 0$ define a random variable $L_k : \{s_{1,0}\}S_1^\omega \to \mathbb{Q}$ by

$$L_k(s_{1,0}s_1s_2\cdots) := lr\big(O(s_{1,0})O(s_1)O(s_2)\cdots O(s_{k-1})\big) \,.$$

Denote by $\mathcal{E}_1$ the expectation with respect to $\mathcal{P}_1$. It was proved in [7, proof of Proposition 6] that $\mathcal{E}_1(L_{k+1} \mid L_k = x) = x$ holds for all $x \in \mathbb{Q}$, i.e., the sequence $L_0, L_1, \ldots$ is a martingale. Unfortunately, the differences $|L_{k+1} - L_k|$ are not bounded, neither are the differences $|\log L_{k+1} - \log L_k|$, as the following example shows.

EXAMPLE 6. *Consider the HMCs H_1, H_2 in Figure 3. For $n > 1$, the probability that H_1 generates the string a^n is $(\frac{1}{3})^{n-1} + \frac{1}{3} \cdot \sum_{i=0}^{n-2}(\frac{1}{3})^i$ which is easily shown to be $\frac{1}{2}(1 + (\frac{1}{3})^{n-1})$, and the probability that H_1 generates a^nb is $(\frac{1}{3})^n$. The corresponding probabilities for H_2 are $(\frac{1}{2})^{n-1}$ and $(\frac{1}{2})^n$, respectively. Now consider any $\alpha \in \{s_0^n s_2\}\{s_0, s_1, s_2\}^\omega$, for some $n > 1$. The two likelihood ratios $L_n(\alpha)$ and $L_{n+1}(\alpha)$ corresponding to the length n and*

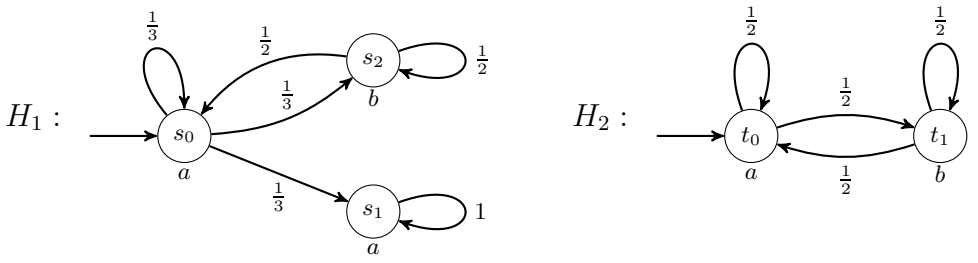

Figure 3: Two HMCs where the difference in log-likelihood ratios is unbounded

length $n+1$ prefixes of α, are given by $L_n(\alpha) = \frac{(\frac{1}{2})^{n-1}}{\frac{1}{2}(1+(\frac{1}{3})^{n-1})}$ and $L_{n+1}(\alpha) = (\frac{3}{2})^n$. Since $n > 1$, we see that $L_n(\alpha) < 2 \cdot (\frac{1}{2})^{n-1} \leq 1$. Hence, $\frac{L_{n+1}(\alpha)}{L_n(\alpha)} > (\frac{3}{2})^n$. So we have that $\log(L_{n+1}(\alpha)) - \log(L_n(\alpha)) > n \cdot \log(\frac{3}{2})$, which is unbounded with increasing n. In a more general case, if α has b appearing infinitely often with an increasing number of a-symbols between two successive b-symbols, then the difference in the log-likelihood ratio of two successive prefixes of α, with the second prefix ending with b, is unbounded. □

This problem of unbounded differences between subsequent log-likelihood ratios prohibits a standard error analysis of hypothesis-testing methods from sequential analysis [30]. Moreover, Azuma's inequality then does not yield an exponentially decaying error bound. As a consequence, we cannot directly prove the bounds of Theorem 5. Therefore, in this subsection, we take a detour. First we develop another monitor M_2' that is not based on likelihoods but is based on a random walk. Then we prove error bounds for M_2'. Then we show that the error bounds for M_2' carry over to the likelihood-based monitor M_2.

The monitor M_2' also runs in N phases, receiving m observations in each phase. The monitor maintains two probability distributions $\psi_1 \in \mathit{Distr}(S_1)$, $\psi_2 \in \mathit{Distr}(S_2)$, and a variable x that takes rational values. Initially, ψ_1, ψ_2 are set to $\delta_{s_{1,0}}, \delta_{s_{2,0}}$ respectively, and x is initialized to 0. The monitor keeps track of $\psi_i = cd_i(\delta_{s_{i,0}}, u)$, for $i = 1, 2$, where u is the observation string received thus far. The variable x indicates a current estimate about which of the two HMCs is being observed: a negative value of x indicates a preference for H_1; a positive value indicates a preference for H_2. In each phase, M_2' waits until it gets the next m observations and then updates x, ψ_1 and ψ_2.

We describe a phase of M_2'. Let ψ_1, ψ_2, x be the values at the end of the previous phase. Let $p_1 = pr_1(\psi_1, \mathcal{A}(\psi_1, \psi_2))$ and $p_2 = pr_2(\psi_2, \mathcal{A}(\psi_1, \psi_2))$. By the definition of a profile we have $p_1 - p_2 \geq c > 0$. Denote by $v \in \Sigma^m$ the string of observations received in the current phase. Assume that $pr_1(\psi_1, v) > 0$ and $pr_2(\psi_2, v) > 0$ (i.e., v can be generated with non-zero probability by both H_1, H_2 from ψ_1, ψ_2 respectively). If $p_1 + p_2 \leq 1$ then x is updated as follows:

$$x := \begin{cases} x - 1 & \text{if } v \in \mathcal{A}(\psi_1, \psi_2) \\ x + \frac{p_1 + p_2}{2 - p_1 - p_2} & \text{if } v \notin \mathcal{A}(\psi_1, \psi_2) \end{cases}$$

If $p_1 + p_2 > 1$ then x is updated as follows:

$$x := \begin{cases} x - \frac{2 - p_1 - p_2}{p_1 + p_2} & \text{if } v \in \mathcal{A}(\psi_1, \psi_2) \\ x + 1 & \text{if } v \notin \mathcal{A}(\psi_1, \psi_2) \end{cases}$$

Note that in all cases, the value of x is increased or decreased by at most 1. After this, ψ_1, ψ_2 are set to $cd_1(\psi_1, v)$ and $cd_2(\psi_2, v)$ respectively, and the phase is finished. On the other hand, if $pr_1(\psi_1, v) > 0$ and $pr_2(\psi_2, v) = 0$ then 1 is output; if $pr_1(\psi_1, v) = 0$ and $pr_2(\psi_2, v) > 0$ then 2 is output; if $pr_1(\psi_1, v) = 0$ and $pr_2(\psi_2, v) = 0$ then 3 is output. In those cases the monitor terminates immediately.

After N phases, if $x \leq 0$ then the monitor M_2' outputs 1, otherwise it outputs 2. An output of i indicates that the sequence is believed to be generated by H_i. Note that M_2'—in contrast to M_2—needs access to the function $\mathcal{A}$. By constructing a supermartingale and applying Azuma's inequality we obtain:

THEOREM 7. *Consider the monitor M_2' running N phases. Let $\mathcal{L}(M_2') \subseteq \Sigma^\omega$ be the set of observation sequences for which M_2' outputs 1. Then,*

$$\mathcal{P}_1([\mathcal{L}(M_2')]_1) \;\geq\; 1 - \exp\left(-\frac{c^2}{18} \cdot N\right) \quad \textit{and}$$
$$\mathcal{P}_2([\mathcal{L}(M_2')]_2) \;\leq\; \exp\left(-\frac{c^2}{18} \cdot N\right).$$

Hence the error probability decays exponentially with N. To prove Theorem 5 we show (in the appendix) that the same error bound, up to a factor of 2, holds for the likelihood-based monitor M_2. The authors are not aware of a proof of Theorem 5 that avoids reasoning about a monitor like M_2'. The proof shows that the difference $\mathcal{P}_1([\mathcal{L}(M_2)]_1) - \mathcal{P}_2([\mathcal{L}(M_2)]_2)$ cannot be increased by any other monitor that is based solely on the first $N \cdot m$ observations: M_2 is optimal in that respect.

To guarantee an error probability bound of at most ε of the likelihood-based monitor M_2, we set $N = \lceil \frac{18}{c^2} \cdot \log\left(\frac{2}{\varepsilon}\right) \rceil$.

EXAMPLE 8. *Figure 4 shows two HMCs H_1, H_2 with a parameter $\delta \in (0, \frac{1}{4}]$. In every step except the first one, H_1 outputs a with probability $\frac{1}{2} + \delta$, and b with probability $\frac{1}{2} - \delta$. For H_2 the probabilities are reversed. The HMCs are distinguishable. The intuitive reason is that H_1 tends to output more a-symbols than b-symbols, whereas H_2 tends to output more b-symbols than a-symbols, and this difference is exhibited in the long run. Intuitively speaking, the smaller δ is, the "less distinguishable" are H_1 and H_2. We will show later that there is a profile with $c = \delta$. By Theorem 5, the probability that M_2 mistakes H_2 for H_1 decays exponentially. More specifically, for an error bound of ε it suffices to make $B\delta^{-2} \log \frac{1}{\varepsilon}$ observations, for a constant $B > 0$. It can be shown that there is a constant $d > 0$ such that M_2 needs, for small ε, at least $d\delta^{-2} \log \frac{1}{\varepsilon}$ observations to push the error*

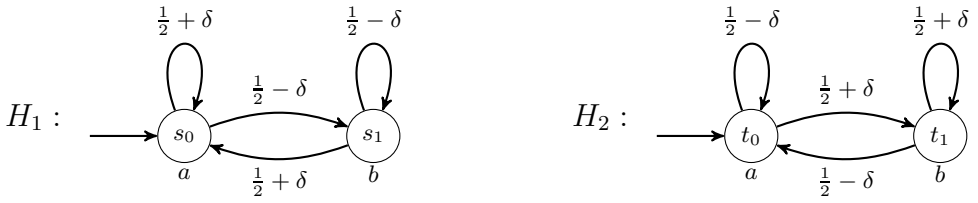

Figure 4: Two distinguishable HMCs with a parameter $\delta \in (0, \frac{1}{4}]$

probability below ε. Hence, for the HMCs from Figure 4, the bound of Theorem 5 is asymptotically tight. As mentioned after the proof of Theorem 5, the likelihood-based monitor is essentially optimal among monitors that observe a fixed-length prefix. So the bound from Theorem 7 is also asymptotically tight. □

4.3 Monitors with One-Sided Error

Now we present M_1, a likelihood-based monitor with one-sided error. Monitor M_1 uses a threshold parameter $low \in (0, 1]$. For each $N > 0$, after reading a prefix v, of length $N \cdot m$, of observations, it computes the likelihood ratio $lr(v)$. If $lr(v) \leq low$, it terminates outputting 1, otherwise it continues.

For any infinite sequence u and integer $i > 0$, let $u[i]$ denote the prefix of u of length i. We fix an integer $N > 0$. Let U_N be the set of all $u \in \Sigma^\omega$ such that $lr(u[N \cdot m]) \leq \exp\big(-\frac{c^2}{36} \cdot N\big)$. Recall from Theorem 5 the set $\mathcal{L}(M_2)$ of observation sequences for which M_2 outputs 1. It should be easy to see that $U_N \subseteq \mathcal{L}(M_2)$. We need the following technical lemma.

LEMMA 9. $\mathcal{P}_1([U_N]_1) \;\geq\; 1 - 4\exp\Big(-\dfrac{c^2}{36} \cdot N\Big)$

This allows us to prove the following theorem:

THEOREM 10. *Consider the monitor M_1 with threshold parameter $low \in (0, 1]$. Let $\mathcal{L}(M_1) \subseteq \Sigma^\omega$ be the set of observation sequences for which M_1 terminates (and hence outputs 1). Then,*

$$\begin{aligned} \mathcal{P}_1([\mathcal{L}(M_1)]_1) &= 1 \qquad \textit{and} \\ \mathcal{P}_2([\mathcal{L}(M_1)]_2) &\leq low\,. \end{aligned}$$

Now we analyze the response time of M_1 taken on observation sequences generated by H_1. Formally, we define a random variable $T : \{s_{1,0}\}S_1^\omega \to \mathbb{N}$ such that T is the number of observations made by monitor M_1 before outputting 1. The following proposition bounds the expected value of T in H_1.

PROPOSITION 11. $\mathcal{E}_1(T) \leq \frac{36m}{c^2} \cdot \log \frac{1}{low} + \frac{147m}{c^2} \cdot low + m$, *where $\mathcal{E}_1(T)$ is the expected value of T under the probability measure $\mathcal{P}_1$.*

The proof of this proposition employs ideas similar to those in [27] for proving an upper bound on the expected monitoring time for exponentially converging monitorable systems. Observe that as *low* decreases, the first term of the bound dominates.

4.4 Monitors for Distinguishing Among Multiple HMCs

Now we address the problem of distinguishing among multiple mutually distinguishable HMCs. We present a monitor based on likelihoods. For $i = 1, \ldots, k$, let $H_i = (G_i, O_i, s_{i,0})$ be HMCs with the same observation alphabet Σ where $G_i = (S_i, R_i, \phi_i)$. Let $\mathcal{P}_i$ and $[\cdot]_i$ be the associated probability measures and inverse observation functions corresponding to the HMC H_i. We assume that they are mutually distinguishable, i.e., for $1 \leq i < j \leq k$, HMCs H_i and H_j are distinguishable. So by Proposition 4 there are profiles $(\mathcal{A}_{i,j}, c_{i,j})$. Define $c := \min\{c_{i,j} \mid 1 \leq i < j \leq k\}$.

Let $m := 2 \cdot \max\{|S_i| \mid 1 \leq i \leq k\}$ and $N > 0$ be an integer parameter. The following monitor M distinguishes among the k HMCs: it takes an observation sequence $u \in \Sigma^{N \cdot m}$ as input and outputs the smallest integer $i \in \{1, \ldots, k\}$ such that $pr_i(\delta_{s_{i,0}}, u) \geq pr_j(\delta_{s_{j,0}}, u)$ for all $j \in \{1, \ldots, k\}$. Essentially, M outputs the index of the HMC whose likelihood value is the highest after $N \cdot m$ observations. By applying the union bound to Theorem 5 we get:

THEOREM 12. *Consider the monitor M. Let $i \in \{1, \ldots, k\}$ and let $\mathcal{L}_i \subseteq \Sigma^{N \cdot m}$ be the set of observation sequences for which M outputs i. Then we have for all $j \in \{1, \ldots, k\} - \{i\}$:*

$$\begin{aligned} \mathcal{P}_i([\mathcal{L}_i\Sigma^\omega]_i) &\;\geq\; 1 - 2k \cdot \exp\left(-\frac{c^2}{18} \cdot N\right) \quad \textit{and} \\ \mathcal{P}_j([\mathcal{L}_i\Sigma^\omega]_j) &\;\leq\; 2\exp\left(-\frac{c^2}{18} \cdot N\right) \end{aligned}$$

5. COMPUTING PROFILES

In the monitors of Section 4 the constant $c > 0$ determines the number N of phases needed to ensure a bound on the error probability. Recall that c is the constant in a profile $(\mathcal{A}, c)$. Any such constant c will do, but the larger it is the better, since the number of phases used will be smaller. Note that even the existence of a positive c (as claimed by Proposition 4) is not obvious. In this section, we prove Theorem 13—which strengthens Proposition 4—by presenting a polynomial-time algorithm to compute a positive c and also the representation of a profile function $\mathcal{A}$ in polynomial time.

Let a *test set* $Test \subseteq \Sigma^*$ be a set of at most m words, with $|v| < m$ for all $v \in Test$. This defines a function $\mathcal{A}_{Test} : Distr(S_1) \times Distr(S_2) \to 2^{\Sigma^m}$ in the following way. Fix $\psi_1 \in Distr(S_1)$ and $\psi_2 \in Distr(S_2)$. Let $v \in Test$ be such that

$$v := \arg \max_{w \in Test} |pr_1(\psi_1, w) - pr_2(\psi_2, w)| \qquad (2)$$

and write

$$[\![v]\!] := \{vw \mid w \in \Sigma^*,\ |vw| = m\}$$

for the set of strings of length m with v as a prefix. Then define:

$$\mathcal{A}_{Test}(\psi_1, \psi_2) := \begin{cases} [\![v]\!] & \text{if } pr_1(\psi_1, v) > pr_2(\psi_2, v) \\ \Sigma^m - [\![v]\!] & \text{otherwise} \end{cases}$$

Depending on the case above, $pr_i(\psi_i, \mathcal{A}_{Test}(\psi_1, \psi_2))$ is either $pr_i(\psi_i, v)$ or $1 - pr_i(\psi_i, v)$. Hence:

$$pr_1(\psi_1, \mathcal{A}_{Test}(\psi_1, \psi_2)) - pr_2(\psi_2, \mathcal{A}_{Test}(\psi_1, \psi_2)) = |pr_1(\psi_1, v) - pr_2(\psi_2, v)| \quad (3)$$

Given a test set *Test* and distributions ψ_1, ψ_2, a monitor can compute the word v from (2) using (1), and hence the probabilities $pr_i(\psi_i, \mathcal{A}_{Test}(\psi_1, \psi_2))$. Moreover, a monitor can check whether a given word $w \in \Sigma^m$ is in $\mathcal{A}_{Test}(\psi_1, \psi_2)$ by checking whether v is a prefix of w.

THEOREM 13. *Let HMCs H_1, H_2 be distinguishable. One can compute, in polynomial time, a test set $Test \subseteq \Sigma^*$ and a number $c > 0$ such that $(\mathcal{A}_{Test}, c)$ is a profile.*

The proof builds on [7] but requires further insights. For the proof we need the concept of *equivalence*: For $i = 1, 2$ let $\psi_i \in Distr(S_i)$. We say that ψ_1 is *equivalent* to ψ_2, written as $\psi_1 \equiv \psi_2$, if $pr_1(\psi_1, u) = pr_2(\psi_2, u)$ holds for all $u \in \Sigma^*$. We have the following proposition:

PROPOSITION 14. *One can compute, in polynomial time, a test set $Test \subseteq \Sigma^*$ such that for all $\psi_1 \in Distr(S_1)$ and all $\psi_2 \in Distr(S_2)$ we have:*

$$\psi_1 \equiv \psi_2 \iff \forall u \in Test:\ pr_1(\psi_1, u) = pr_2(\psi_2, u)$$

The algorithm for Proposition 14 uses linear-algebra based techniques that have been developed for deciding equivalence of HMCs, see e.g. [25, 29, 15, 10].

We fix *Test* for the remainder of the section. We define a distance measure $dist(\psi_1, \psi_2)$ between ψ_1, ψ_2 given by

$$dist(\psi_1, \psi_2) := \max_{w \in Test} |pr_1(\psi_1, w) - pr_2(\psi_2, w)|.$$

By Proposition 14 we have:

$$\psi_1 \equiv \psi_2 \iff dist(\psi_1, \psi_2) = 0$$

For the following proposition, linear programming is used to compute a lower bound on $dist(\psi_1, \psi_2)$ for reachable pairs (ψ_1, ψ_2) in distinguishable HMCs:

PROPOSITION 15. *Let H_1, H_2 be distinguishable HMCs. One can compute, in polynomial time, a rational number $c > 0$ such that for all reachable pairs (ψ_1, ψ_2) of distributions we have $dist(\psi_1, \psi_2) \geq c$.*

In general there may exist *unreachable* pairs (ψ_1, ψ_2) of distributions with $dist(\psi_1, \psi_2) = 0$, even for distinguishable HMCs. Proposition 15 establishes in particular the nontrivial fact that for distinguishable HMCs there *exists* a positive lower bound on $dist(\psi_1, \psi_2)$ for all reachable pairs (ψ_1, ψ_2).

EXAMPLE 16. *Consider again the HMCs from Figure 4. We compute the set Test according to the algorithm from Proposition 14. This yields $Test = \{\varepsilon, a, aa, ba\}$, where ε denotes the empty word.*

In this example, the last symbol of any observation sequence reveals the state. Hence there are only two reachable pairs of distributions: one is (π_1, π_2) with $\pi_1(s_0) = \pi_2(t_0) = 1$, and the other one is (π_1', π_2') with $\pi_1'(s_1) = \pi_2'(t_1) = 1$. Using the definition of dist we compute:

$$\begin{aligned} dist(\pi_1, \pi_2) &= pr_1(\pi_1, aa) - pr_2(\pi_2, aa) \\ &= \frac{1}{2} + \delta - \left(\frac{1}{2} - \delta\right) = 2\delta \\ dist(\pi_1', \pi_2') &= pr_1(\pi_1', ba) - pr_2(\pi_2', ba) \\ &= \frac{1}{2} + \delta - \left(\frac{1}{2} - \delta\right) = 2\delta \end{aligned}$$

Hence we have $dist(\psi_1, \psi_2) = 2\delta > 0$ for all reachable *pairs (ψ_1, ψ_2) of distributions.*

In order to illustrate some aspects of Proposition 15, we use linear programming to compute a lower bound on $dist(\psi_1, \psi_2)$ for all *(reachable or unreachable) pairs (ψ_1, ψ_2) of distributions. Concretely, we solve the following linear program, where δ is the constant parameter from the HMCs H_1, H_2, and the variables are x and variables encoding distributions ψ_1, ψ_2:*

minimize $x \geq 0$

subject to: $\psi_1 \in Distr(S_1),\ \psi_2 \in Distr(S_2)$,

$$-x \leq pr_1(\psi_1, u) - pr_2(\psi_2, u) \leq x \text{ for all } u \in Test.$$

An optimal solution is $x = \delta$ and $\psi_1(s_0) = \frac{3}{4} - \frac{\delta}{2}$ and $\psi_1(s_1) = \frac{1}{4} + \frac{\delta}{2}$ and $\psi_2(t_0) = \frac{3}{4} + \frac{\delta}{2}$ and $\psi_2(t_1) = \frac{1}{4} - \frac{\delta}{2}$. Hence $x = \delta > 0$ is a lower bound on $dist(\psi_1, \psi_2)$ for all *pairs of distributions, and hence, a fortiori, also for all reachable pairs. As mentioned after Proposition 15, the reachability aspect is in general (unlike in this example) essential for obtaining a positive lower bound. Indeed, the proof of Proposition 15 takes advantage of further results from [7].*

If we compute a lower bound according to the proof Proposition 15, i.e., taking reachability into account, we obtain $c = 4\delta/(3 + 2\delta)$, which lies strictly between the previously computed lower bounds δ and 2δ. □

With Proposition 15 at hand, we are ready to prove Theorem 13:

PROOF OF THEOREM 13. Compute *Test* according to Proposition 14 and $c > 0$ according to Proposition 15. We show that $(\mathcal{A}_{Test}, c)$ is a profile. Let (ψ_1, ψ_2) be a reachable pair of distributions. Let $v \in \Sigma^*$ be as in (2). We have:

$$\begin{aligned} &pr_1(\psi_1, \mathcal{A}_{Test}(\psi_1, \psi_2)) - pr_2(\psi_2, \mathcal{A}_{Test}(\psi_1, \psi_2)) \\ &= |pr_1(\psi_1, v) - pr_2(\psi_2, v)| && \text{by (3)} \\ &= \max_{w \in Test} |pr_1(\psi_1, w) - pr_2(\psi_2, w)| && \text{by (2)} \\ &= dist(\psi_1, \psi_2) && \text{def. of } dist \\ &\geq c && \text{Proposition 15} \end{aligned}$$

This completes the proof. □

We have seen that for a given error bound, the number of observations our monitors need to make depends quadratically on $\frac{1}{c}$. So it may be beneficial to compute a larger value of c, even if such a computation is expensive. To this end, for a distribution $\pi \in Distr(S)$, write $supp(\pi) := \{s \in S \mid \pi(s) > 0\}$. For HMCs H_1, H_2, if a

pair (ψ_1, ψ_2) of distributions is reachable, we say that the pair $(supp(\psi_1), supp(\psi_2))$ is *reachable*. We have the following proposition:

PROPOSITION 17. *Let H_1, H_2 be two distinguishable HMCs. One can compute, in exponential time:*

$$c := \min_{reachable\ (S_1', S_2') \in 2^{S_1} \times 2^{S_2}} \ \min_{\psi_1 \in Distr(S_1')} \ \min_{\psi_2 \in Distr(S_2')} \max_{U \subseteq \Sigma^m} \left(pr_1(\psi_1, U) - pr_2(\psi_2, U) \right)$$

(Note that U ranges over a set of double-exponential size.) This value of c is lower-bounded by the value of $c > 0$ from Theorem 13, and it is part of a profile with

$$\mathcal{A}(\psi_1, \psi_2) \ = \ \arg\max_{U \subseteq \Sigma^m} \left(pr_1(\psi_1, U) - pr_2(\psi_2, U) \right).$$

6. APPLICATION: RUNTIME VERIFICATION

In this section we discuss an application of monitors for runtime verification of stochastic systems. Traditional verification aims at proving correctness of systems at the time of their design. This quickly becomes infeasible, in particular for complex systems with several components and stochastic behavior, see e.g. [26]. *Runtime verification* is an alternative where a monitor observes a system while it is running, and raises an alarm once a faulty behavior is detected. The alarm may trigger, e.g., a fail-safe way of shutting the system down. HMCs were suggested in [26, 27] as models of partially observable stochastic systems. In this section, the monitor does not try to distinguish two HMCs, rather it tries to distinguish correct and faulty behavior of a single HMC.

Definitions. For a probability measure $\mathcal{P}$ and measurable sets C, D such that $\mathcal{P}(C) > 0$, we let $\mathcal{P}(D \mid C)$ denote the value $\frac{\mathcal{P}(C \cap D)}{\mathcal{P}(C)}$, which is the conditional probability of D given C. A *classifying HMC (cHMC)* is a quadruple $H = (G, O, s_0, Class)$, where (G, O, s_0) is an HMC and *Class* is a condition classifying each bottom strongly connected component (BSCC) of H as *bad* or *good*. For a cHMC and a state $s \in S$ we define:

$$Bad_s := \{ss_1 s_2 \cdots \in \{s\}S^\omega \mid \exists i : s_i \text{ is in a bad BSCC}\}$$
$$Good_s := \{ss_1 s_2 \cdots \in \{s\}S^\omega \mid \exists i : s_i \text{ is in a good BSCC}\}$$

Define $Bad := Bad_{s_0}$ and $Good := Good_{s_0}$. The events *Bad* and *Good* are disjoint and measurable. By fundamental properties of Markov chains we have

$$\mathcal{P}(Bad \cup Good) = \mathcal{P}(Bad) + \mathcal{P}(Good) = 1\,.$$

To avoid trivialities we assume that $\mathcal{P}(Bad), \mathcal{P}(Good) > 0$ (this can be checked in polynomial time by graph reachability). We say that a cHMC H is *monitorable* if for every $\varepsilon > 0$ there exists a monitor M such that

$$\mathcal{P}([\mathcal{L}(M)] \mid Bad) \ \geq \ 1 - \varepsilon \quad \text{and}$$
$$\mathcal{P}([\mathcal{L}(M)] \mid Good) \ \leq \ \varepsilon\,.$$

In [26] the authors define and study monitorability of pairs $(H_0, \mathcal{A})$ where H_0 is an HMC and $\mathcal{A}$ is a deterministic Streett automaton. One can compute, in polynomial time, the product of H_0 and $\mathcal{A}$. That product is a cHMC H as defined above. Then $(H_0, \mathcal{A})$ is monitorable (in the sense of [26]) if and only if H is monitorable (in the sense defined above).

A construction similar to one that was given in [4, Section 3] allows us, for a given cHMC H, to construct two HMCs H_1, H_2 that exhibit the bad and the good behavior of H according to their conditional probabilities:

PROPOSITION 18. *Let H be a cHMC with $\mathcal{P}(Bad), \mathcal{P}(Good) > 0$. Then one can compute, in polynomial time, HMCs H_1, H_2 such that for all measurable events $E \subseteq S^\omega$ we have*

$$\mathcal{P}_1(E) = \mathcal{P}(E \mid Bad) \quad \textit{and} \quad \mathcal{P}_2(E) = \mathcal{P}(E \mid Good)\,.$$

It follows from Proposition 18 that distinguishing and monitoring are equivalent: Given HMCs H_1, H_2, we can combine them into a single cHMC H by introducing a new initial state s_0, which branches to the initial states of H_1, H_2 with probability 1/2 each. We classify the BSCCs of H_1 and of H_2 as bad and good, respectively. Then for any $E \subseteq \Sigma^\omega$ we have

$$\mathcal{P}_1(E) = \mathcal{P}(\{O(s_0)\}E \mid Bad) \quad \text{and}$$
$$\mathcal{P}_2(E) = \mathcal{P}(\{O(s_0)\}E \mid Good)\,,$$

so any monitor for H can be translated in a straightforward way into a monitor that distinguishes H_1 and H_2. Conversely, given a cHMC H, we can compute H_1, H_2 according to Proposition 18. Then any monitor that distinguishes H_1 and H_2 also monitors H.

By combining this observation with Theorem 3 we obtain:

COROLLARY 19. *One can decide in polynomial time whether a given cHMC H is monitorable.*

Another kind of monitorability, called *strong monitorability* [26], was shown PSPACE-complete in [26]. Strong monitorability implies monitorability.

Using Proposition 18 again, the monitors from Section 4 apply to monitoring cHMCs. For instance, the monitor with one-sided error can guarantee that (a) given that the behavior is faulty then an alarm is raised with probability 1 and within short expected time, and (b) given that the behavior is correct then probably no alarm is raised.

7. CONCLUSIONS

In this paper we have considered the distinguishability problem for HMCs. We have shown that it is decidable in polynomial time.

We have presented two likelihood based monitors M_1, M_2 for distinguishing between HMCs H_1, H_2 based on the sequences of observations generated by them. The monitor M_2 makes a decision after running for a fixed number of observations and exhibits two-sided error. It processes $O(\log \frac{1}{\varepsilon})$ observations to ensure an error probability of at most ε. The monitor M_1 has only one-sided error. The expected number of observations it processes to identify a sequence generated by H_1 is $O(\log \frac{1}{\varepsilon})$ to guarantee an error probability of at most ε on sequences generated by H_2. We have also provided a monitor for distinguishing multiple HMCs. All error analyses rely on martingale techniques, in particular, Azuma's inequality.

Polynomial time bounded algorithms are provided, which for the monitor M_2, compute the number of observations that guarantees a given upper bound on the error, and for the M_1 compute the expected number of observations of H_1

before which an alarm is raised, for a given error bound on the probability of raising an alarm on inputs generated by H_2. These algorithms employ linear programming based techniques for computing profiles.

We have discussed an application to runtime verification of stochastic systems. The monitorability problem for cHMCs is polynomial-time equivalent to distinguishability, and hence decidable in polynomial time. We have shown that the monitors developed in this paper can be adapted so that they monitor cHMCs.

One direction for future work is to improve the efficiency of computing a good lower bound on c. We have seen that this bound strongly influences the number of observations the monitor needs to make, so the bound may determine the applicability of a monitor in practice. Another direction is to develop a notion of a monitor for HMCs that are not equivalent but not distinguishable. Such monitors might still attempt to distinguish between the HMCs for as many runs as possible.

Acknowledgments. Stefan Kiefer is supported by a University Research Fellowship of the Royal Society. Prasad Sistla is partly supported by the NSF grants CCF-1319754 and CNS-1314485.

8. 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] 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.

[3] R. Alur, C. Courcoubetis, and M. Yannakakis. Distinguishing tests for nondeterministic and probabilistic machines. In *Proceedings of STOC*, pages 363–372. ACM, 1995.

[4] C. Baier, J. Klein, S. Klüppelholz, and S. Märcker. Computing conditional probabilities in Markovian models efficiently. In *Proceedings of TACAS*, volume 8413 of *LNCS*, pages 515–530, 2014.

[5] N. Bertrand, S. Haddad, and E. Lefaucheux. Accurate approximate diagnosability of stochastic systems. In *Proceedings of LATA*, pages 549–561, 2016.

[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 CSL-LICS*, pages 33:1–33:10, 2014.

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

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

[10] L. Doyen, T. Henzinger, and J.-F. Raskin. Equivalence of labeled Markov chains. *International Journal of Foundations of Computer Science*, 19(3):549–563, 2008.

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

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

[13] A. Fraser. *Hidden Markov Models and Dynamical Systems*. Society for Industrial and Applied Mathematics, 2008.

[14] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4):903–919, 2001.

[15] H. Ito, S.-I. Amari, and K. Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Transactions on Information Theory*, 38(2):324–333, March 1992.

[16] B.-H. Juang and L. Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64(2):391–408, February 1985.

[17] S. Kiefer and A. Sistla. Distinguishing hidden Markov chains. Technical report, arxiv.org, 2015. Available at `http://arxiv.org/abs/1507.02314`.

[18] A. Krogh, B. Larsson, G. von Heijne, and E. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, 305(3):567–580, 2001.

[19] R. Lyngsø and C. Pedersen. The consensus string problem and the complexity of comparing hidden Markov models. *Journal of Computer and System Sciences*, 65(3):545–569, 2002.

[20] A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*. McGrawHill, NewYork, 2002.

[21] R. Phatarfod. Sequential analysis of dependent observations. I. *Biometrika*, 52(1-2):157–165, 1965.

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

[23] C. Raphael. Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):360–370, April 1999.

[24] N. Schmitz and B. Süselbeck. Sequential probability ratio tests for homogeneous Markov chains. In *Mathematical Learning Models – Theory and Algorithms*, volume 20 of *Lecture Notes in Statistics*, pages 191–202, 1983.

[25] M.-P. Schützenberger. On the definition of a family of automata. *Inf. and Control*, 4:245–270, 1961.

[26] A. Sistla, M. Žefran, and Y. Feng. Monitorability of stochastic dynamical systems. In *Proceedings of CAV*, volume 6806 of *LNCS*, pages 720–736, 2011.

[27] A. Sistla, M. Žefran, Y. Feng, and Y. Ben. Timely monitoring of partially observable stochastic systems. In *Proceedings of the 17th international conference on Hybrid systems: computation and control (HSCC14)*, pages 61–70, 2014.

[28] R. Swamy. Sequential comparison of two Markov chains. *Biometrika*, 70(1):293–296, 1983.

[29] W. Tzeng. A polynomial-time algorithm for the equivalence of probabilistic automata. *SIAM Journal on Computing*, 21(2):216–227, 1992.

[30] B. Wetherill and K. Glazenbrook. *Sequential Methods in Statistics*. Chapman and Hall, 1986.

[31] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

APPENDIX

A. PROOFS OF SECTION 4

We prove Theorem 7 from the main text.

THEOREM 7. *Consider the monitor M_2' running N phases. Let $\mathcal{L}(M_2') \subseteq \Sigma^\omega$ be the set of observation sequences for which M_2' outputs 1. Then,*

$$\mathcal{P}_1([\mathcal{L}(M_2')]_1) \;\geq\; 1 - \exp\left(-\frac{c^2}{18}\cdot N\right) \quad \textit{and}$$
$$\mathcal{P}_2([\mathcal{L}(M_2')]_2) \;\leq\; \exp\left(-\frac{c^2}{18}\cdot N\right).$$

PROOF. Let $\psi_{1,k}, \psi_{2,k}$ and X_k denote the values of ψ_1, ψ_2 and x directly after the k-th phase, for $k \geq 0$. Initially, $\psi_{1,0} = \delta_{s_{1,0}}$, $\psi_{2,0} = \delta_{s_{2,0}}$ and $X_0 = 0$. For $k \geq 1$, let $u_k \in \Sigma^{k\cdot m}$ be the sequence of all observations received until and including phase k. By induction on k, it is easy to see that $\psi_{i,k} = cd_i(\delta_{s_{i,0}}, u_k)$ for $k \geq 0$, $i = 1, 2$.

Note that X_k depends only on u_k. In the following we view $X_0, X_1, \ldots$ as a sequence of random variables. (Formally, for $k \geq 0$ the random variable X_k is a function of type $X_k : \{s_{1,0}\}S_1^\omega \to \mathbb{Q}$.) We also define a sequence $Y_0, Y_1, \ldots$ of random variables with $Y_k = X_k + k \cdot \frac{c}{2}$. Note that $X_0 = Y_0 = 0$.

We show that the sequence of random variables $Y_0, Y_1, \ldots$ forms a supermartingale in H_1. Let $k \geq 0$. Fix $u_k \in \Sigma^{k\cdot m}$. Recall that this determines X_k. For the conditional expected value of X_{k+1} given prefix u_k we have:

$$\mathcal{E}(X_{k+1} \mid [u_k\Sigma^\omega]_1) \;=\; X_k + d\,, \tag{4}$$

where d denotes the expected change of x after phase $k+1$. Recall that $\psi_{i,k} = cd_i(\delta_{s_{i,0}}, u_k)$ for $i = 1, 2$. Let $p_i = pr_i(\psi_{i,k}, \mathcal{A}(\psi_{1,k}, \psi_{2,k}))$, for $i = 1, 2$. Assume $p_1 + p_2 \leq 1$. According to our rule for updating x we then have:

$$d \;=\; p_1 \cdot (-1) + (1 - p_1) \cdot \frac{p_1 + p_2}{2 - p_1 - p_2} \;=\; \frac{p_2 - p_1}{2 - p_1 - p_2}$$

This is negative. Moreover, by the definition of a profile we have $p_1 - p_2 \geq c > 0$. Further more, $1 \leq 2 - p_1 - p_2 < 2$. Hence:

$$d \;\leq\; \frac{p_2 - p_1}{2} \;\leq\; -\frac{c}{2}$$

Combining this with (4) and the definition of Y_k we obtain:

$$\mathcal{E}(Y_{k+1} \mid [u_k\Sigma^\omega]_1) \;=\; Y_k + \frac{c}{2} + d \;\leq\; Y_k \tag{5}$$

Now assume $p_1 + p_2 > 1$. Then we have:

$$\begin{aligned} d \;&=\; -p_1 \cdot \frac{2 - p_1 - p_2}{p_1 + p_2} + (1 - p_1) \cdot 1 \;=\; \frac{p_2 - p_1}{p_1 + p_2} \\ &\leq\; \frac{p_2 - p_1}{2} \;\leq\; -\frac{c}{2}\,, \end{aligned}$$

so (5) again follows. Hence we have shown that $Y_0, Y_1, \ldots$ is a supermartingale in H_1.

By definition of the update rule we have $|X_{k+1} - X_k| \leq 1$ and hence $|Y_{k+1} - Y_k| \leq 1 + \frac{c}{2} \leq \frac{3}{2}$. Applying Azuma's inequality (see, e.g., [31]) to the supermartingale $Y_0, Y_1, \ldots$ we obtain:

$$\begin{aligned} \mathcal{P}_1\{X_N > 0\} \;&=\; \mathcal{P}_1\left\{Y_N > \frac{c}{2}\cdot N\right\} \;\leq\; \exp\left(-\frac{\left(\frac{c}{2}\cdot N\right)^2}{2N\cdot\left(\frac{3}{2}\right)^2}\right) \\ &=\; \exp\left(-\frac{c^2}{18}\cdot N\right) \end{aligned}$$

Hence,

$$\mathcal{P}_1([\mathcal{L}(M)]_1) \;=\; \mathcal{P}_1\{X_N \leq 0\} \;\geq\; 1 - \exp\left(-\frac{c^2}{18}\cdot N\right).$$

From this, it follows that $\mathcal{P}_1([\mathcal{L}(M)]_1) \geq 1 - \exp(-\frac{c^2}{18}\cdot N)$.

The proof of the second inequality in the statement is similar with the following modifications. The random variables X_k' are defined like X_k, but on sequences of states in H_2 rather than H_1. Define $Y_k' = X_k' - k \cdot \frac{c}{2}$. The sequence $Y_0', Y_1', \ldots$ is now a submartingale. Applying Azuma's inequality to this submartingale now leads to the second inequality claimed in the statement. □

The following lemma is used for the proof of Theorem 5.

LEMMA 20. *Let S be a countable set. Let ψ_1, ψ_2 be probability distributions over S. For $i \in \{1, 2\}$ and any event $V \subseteq S$ define $\psi_i(V) := \sum_{v\in V} \psi_i(v)$. Define*

$$W := \{s \in S \mid \psi_1(s) \geq \psi_2(s)\}\,.$$

Then

$$\max_{V\subseteq S}\big(\psi_1(V) - \psi_2(V)\big) \;=\; \psi_1(W) - \psi_2(W)\,,$$

i.e., W maximizes the probability difference over all events.

PROOF. If $s \in S$ with $s \notin V$ and $\psi_1(s) \geq \psi_2(s)$, then

$$\begin{aligned} &\psi_1(V \cup \{s\}) - \psi_2(V \cup \{s\}) \\ &= \psi_1(V) - \psi_2(V) + \psi_1(s) - \psi_2(s) \\ &\geq \psi_1(V) - \psi_2(V)\,. \end{aligned}$$

Similarly, if $s \in S$ with $s \in V$ and $\psi_1(s) < \psi_2(s)$, then

$$\begin{aligned} &\psi_1(V \setminus \{s\}) - \psi_2(V \setminus \{s\}) \\ &= \psi_1(V) - \psi_2(V) - \psi_1(s) + \psi_2(s) \\ &> \psi_1(V) - \psi_2(V)\,. \end{aligned}$$

The statement of the lemma follows. □

Now we prove Theorem 5 from the main text.

THEOREM 5. *Consider the monitor M_2 that reads the first $N \cdot m$ observations. Let $\mathcal{L}(M_2) \subseteq \Sigma^\omega$ be the set of observation sequences for which M_2 outputs 1. Then we have*

$$\mathcal{P}_1([\mathcal{L}(M_2)]_1) - \mathcal{P}_2([\mathcal{L}(M_2)]_2) \geq 1 - 2\exp\left(-\frac{c^2}{18}\cdot N\right).$$

Hence,

$$\mathcal{P}_1([\mathcal{L}(M_2)]_1) \;\geq\; 1 - 2\exp\left(-\frac{c^2}{18}\cdot N\right) \quad \textit{and}$$
$$\mathcal{P}_2([\mathcal{L}(M_2)]_2) \;\leq\; 2\exp\left(-\frac{c^2}{18}\cdot N\right).$$

PROOF. Let $N \geq 0$. We can write $\mathcal{L}(M_2) = W\Sigma^\omega$ where $W \subseteq \Sigma^{N\cdot m}$ denotes the set of observation prefixes of length $N \cdot m$ on which M_2 outputs 1. Then we have:

$$\begin{aligned} W &= \{u \in \Sigma^{N\cdot m} \mid pr_1(\delta_{s_{1,0}}, u) \geq pr_2(\delta_{s_{2,0}}, u)\} \\ &= \{u \in \Sigma^{N\cdot m} \mid \mathcal{P}_1([\{u\}\Sigma^\omega]_1) \geq \mathcal{P}_2([\{u\}\Sigma^\omega]_2)\} \end{aligned}$$

(We left the output of the monitor unspecified when the likelihood ratio is equal to 1. As a consequence, the inequalities above might be strict. This does not affect the rest of the argument.) Using Lemma 20 we obtain the following inequality.

$$\mathcal{P}_1([W\Sigma^\omega]_1) - \mathcal{P}_2([W\Sigma^\omega]_2) \;\geq\; \mathcal{P}_1([V\Sigma^\omega]_1) - \mathcal{P}_2([V\Sigma^\omega]_2) \tag{6}$$

for all $V \subseteq \Sigma^{N\cdot m}$. In particular, this holds for the prefixes of length $N \cdot m$ of $\mathcal{L}(M_2')$ from Theorem 7. Hence we have:

$$\begin{aligned} & \mathcal{P}_1([\mathcal{L}(M_2)]_1) - \mathcal{P}_2([\mathcal{L}(M_2)]_2) \\ = \;& \mathcal{P}_1([W\Sigma^\omega]_1) - \mathcal{P}_2([W\Sigma^\omega]_2) && \text{as } \mathcal{L}(M_2) = W\Sigma^\omega \\ \geq \;& \mathcal{P}_1([\mathcal{L}(M_2')]_1) - \mathcal{P}_2([\mathcal{L}(M_2')]_2) && \text{by (6)} \\ \geq \;& 1 - 2\exp\Big(-\frac{c^2}{18}\cdot N\Big) && \text{by Theorem 7} \end{aligned}$$

This concludes the proof of the Theorem 5. □

Computation for Example 8. We analyse the likelihood-based monitor M_2 for the HMCs of Figure 4. The monitor M_2 makes $N \cdot m = 4N$ observations. It is easy to see that it outputs 1 if and only if it reads at least as many a-symbols as b-symbols, i.e., the number of read a-symbols is at least $2N$. Hence we have:

$$\begin{aligned} & \mathcal{P}_2([\mathcal{L}(M_2)]_2) \\ = \;& \sum_{i=2N}^{4N} \binom{4N}{i} \Big(\frac{1}{2}-\delta\Big)^i \Big(\frac{1}{2}+\delta\Big)^{4N-i} \\ \geq \;& \binom{4N}{2N} \Big(\frac{1}{2}-\delta\Big)^{2N} \Big(\frac{1}{2}+\delta\Big)^{2N} \\ = \;& \frac{(4N)!}{(2N)!\cdot(2N)!} \Big(\frac{1}{4}-\delta^2\Big)^{2N} \\ = \;& \frac{2^{2N}\cdot(4N-1)\cdot(4N-3)\cdots 5\cdot 3}{(2N)\cdot(2N-1)\cdot(2N-2)\cdots 2\cdot 1} \Big(\frac{1}{4}-\delta^2\Big)^{2N} \\ \geq \;& \frac{2^{2N}}{(2N)}\cdot 2^{2N-1}\cdot\Big(\frac{1}{4}-\delta^2\Big)^{2N} \\ = \;& \frac{1}{4N}\cdot\big(1-4\delta^2\big)^{2N} \end{aligned}$$

For $x \in [0, \frac{1}{2}]$ we have $\ln(1-x) \geq -2x$. So we can continue as follows:

$$\begin{aligned} & \mathcal{P}_2([\mathcal{L}(M_2)]_2) \\ \geq \;& \frac{1}{4N}\cdot\exp\big(-16\delta^2 N\big) \\ \geq \;& \exp\big(-17\delta^2 N\big) \qquad \text{for large } N \end{aligned}$$

It follows that for small ε, an inequality $\varepsilon \geq \mathcal{P}_2([\mathcal{L}(M_2)]_2)$ implies that $N \geq \frac{1}{17}\delta^{-2}\ln\frac{1}{\varepsilon}$. This completes the calculation for the example. □

We prove Lemma 9 from the main text.

LEMMA 9. $\mathcal{P}_1([U_N]_1) \;\geq\; 1 - 4\exp\Big(-\dfrac{c^2}{36}\cdot N\Big)$

PROOF. By contradiction. Contrary to the lemma, assume:

$$\mathcal{P}_1([U_N]_1) < 1 - 4\exp\Big(-\frac{c^2}{36}\cdot N\Big) \tag{7}$$

Let $V_N := \mathcal{L}(M_2) - U_N$, and let W_N denote the set of all prefixes, of length $N \cdot m$, of sequences in V_N. Clearly, for all $v \in W_N$:

$$\exp\Big(-\frac{c^2}{36}\cdot N\Big) < lr(v) < 1$$

It follows for all $v \in W_N$:

$$pr_1(\delta_{s_{1,0}}, v) = \frac{pr_2(\delta_{s_{2,0}}, v)}{lr(v)} < \exp\Big(\frac{c^2}{36}\cdot N\Big)\cdot pr_2(\delta_{s_{2,0}}, v)$$

$$\begin{aligned} \text{Hence, } \mathcal{P}_1([V_N]_1) &= \sum_{v\in W_N} pr_1(\delta_{s_{1,0}}, v) \\ &< \exp\Big(\frac{c^2}{36}\cdot N\Big)\cdot \underbrace{\sum_{v\in W_N} pr_2(\delta_{s_{2,0}}, v)}_{=\mathcal{P}_2([V_N]_2)} . \end{aligned} \tag{8}$$

From Theorem 5 we know that

$$\mathcal{P}_2([V_N]_2) \;\leq\; \mathcal{P}_2([\mathcal{L}(M_2)]_2) \;\leq\; 2\exp\Big(-\frac{c^2}{18}\cdot N\Big).$$

By combining this with (8), we get:

$$\begin{aligned} \mathcal{P}_1([V_N]_1) &< 2\exp\Big(\frac{c^2}{36}\cdot N\Big)\cdot\exp\Big(-\frac{c^2}{18}\cdot N\Big) \\ &= 2\exp\Big(-\frac{c^2}{36}\cdot N\Big) \end{aligned} \tag{9}$$

We have $\mathcal{P}_1([\mathcal{L}(M_2)]_1) = \mathcal{P}_1([V_N]_1) + \mathcal{P}_1([U_N]_1)$. Using (7) and (9), we get:

$$\mathcal{P}_1([\mathcal{L}(M_2)]_1) < 1 - 2\exp\Big(-\frac{c^2}{36}\cdot N\Big) < 1 - 2\exp\Big(-\frac{c^2}{18}\cdot N\Big)$$

But this contradicts Theorem 5. □

We prove Theorem 10 from the main text:

THEOREM 10. *Consider the monitor M_1 with threshold parameter $low \in (0,1]$. Let $\mathcal{L}(M_1) \subseteq \Sigma^\omega$ be the set of observation sequences for which M_1 terminates (and hence outputs 1). Then,*

$$\begin{aligned} \mathcal{P}_1([\mathcal{L}(M_1)]_1) &= 1 \qquad \textit{and} \\ \mathcal{P}_2([\mathcal{L}(M_1)]_2) &\leq \textit{low}\,. \end{aligned}$$

PROOF. Let N_0 be the smallest integer such that $\exp\Big(-\frac{c^2}{36}\cdot N_0\Big) \leq low$. Clearly, for all $N \geq N_0$ we have $\mathcal{L}(M_1) \supseteq U_N$ where U_N is the set defined at the beginning of Section 4.3. From this observation and Lemma 9, we see that for all $N \geq N_0$:

$$\mathcal{P}_1([\mathcal{L}(M_1)]_1) \geq 1 - 4\exp\Big(-\frac{c^2}{36}\cdot N\Big)$$

From this, we get

$$\mathcal{P}_1([\mathcal{L}(M_1)]_1) \geq \lim_{N\to\infty} 1 - 4\exp\Big(-\frac{c^2}{36}\cdot N\Big) = 1.$$

Let $X = \{v \in (\Sigma^m)^* \mid pr_1(\delta_{s_{1,0}}, v) > 0,\ lr(v) \leq low,\ \forall i < |v| : lr(v[i]) > low\}$. Intuitively, X is the set of shortest observation sequences whose length is a multiple of m and whose likelihood ratio is $\leq low$. It is easy to see that $\mathcal{L}(M_1) = X\Sigma^\omega$. Observe that there do not exist two distinct sequences $v_1, v_2 \in X$ such that v_1 is a prefix of v_2.

$$\begin{aligned}\mathcal{P}_2([\mathcal{L}(M_1)]_2) &= \sum_{v\in X} pr_2(\delta_{s_{2,0}}, v)\\ &\leq low \cdot \sum_{v\in X} pr_1(\delta_{s_{1,0}}, v)\\ &= low \cdot \mathcal{P}_1([\mathcal{L}(M_1)]_1) \leq low\end{aligned}$$

□

We prove Proposition 11 from the main text.

PROPOSITION 11. *$\mathcal{E}_1(T) \leq \frac{36m}{c^2} \cdot \log \frac{1}{low} + \frac{147m}{c^2} \cdot low + m$, where $\mathcal{E}_1(T)$ is the expected value of T under the probability measure $\mathcal{P}_1$.*

PROOF. Since T is a nonnegative integer valued random variable, from [20], we see that $\mathcal{E}_1(T) = \sum_{n\geq 0} \mathcal{P}_1\{T > n\}$. Since M_1 only decides after each phase, i.e., after reading each successive sequence of m observations, we see that $\mathcal{E}_1(T) = \sum_{N\geq 0} m \cdot \mathcal{P}_1\{T > N \cdot m\}$. Let N_0 be the smallest integer such that $\exp\left(-\frac{c^2}{36} \cdot N_0\right) \leq low$, i.e., $N_0 = \lceil \frac{36}{c^2} \cdot \log \frac{1}{low} \rceil$. We have:

$$\begin{aligned}\mathcal{E}_1(T) &= \sum_{N=0}^{N_0-1} m \cdot \underbrace{\mathcal{P}_1\{T > N\cdot m\}}_{\leq 1} + \sum_{N\geq N_0} m \cdot \mathcal{P}_1\{T > N\cdot m\}\\ &\leq m \cdot N_0 + m \cdot \sum_{N\geq N_0} \mathcal{P}_1\{T > N\cdot m\} \qquad (10)\end{aligned}$$

For $N \geq 0$, let $X_N = \{u \in \Sigma^\omega \mid lr(u[N\cdot m]) > low\}$. Observe that, for $N \geq N_0$, $X_N \subseteq \Sigma^\omega - U_N$. Further,

$$\begin{aligned}\mathcal{P}_1\{T > N\cdot m\} &\leq \mathcal{P}_1([X_N]_1) \leq \mathcal{P}_1([\Sigma^\omega - U_N]_1)\\ &\leq 4\exp\left(-\frac{c^2}{36} \cdot N\right) \quad \text{from Lemma 9.}\end{aligned} \qquad (11)$$

From (10) and (11), we get

$$\begin{aligned}\mathcal{E}_1(T) &= m\cdot N_0 + 4m \cdot \sum_{N\geq N_0} \exp\left(-\frac{c^2}{36}\cdot N\right)\\ &= m\cdot N_0 + 4m \cdot \sum_{N\geq 0} \exp\left(-\frac{c^2}{36}\cdot (N+N_0)\right)\\ &= m\cdot N_0 + 4m \cdot \exp\left(-\frac{c^2}{36}\cdot N_0\right) \cdot \sum_{N\geq 0} \exp\left(-\frac{c^2}{36}\cdot N\right)\\ &\leq m\cdot N_0 + 4m\cdot low \cdot \frac{1}{1-\exp\left(-\frac{c^2}{36}\right)}\\ &\leq \frac{36m}{c^2}\cdot \log\frac{1}{low} + m + 4m\cdot low \cdot \frac{1}{1-\exp\left(-\frac{c^2}{36}\right)}\end{aligned} \qquad (12)$$

substituting for N_0.

By using a Taylor series expansion of $\exp\left(-\frac{c^2}{36}\right)$, we get an infinite sum in which the signs of the terms alternate starting with a positive sign, and in which the absolute values of the terms decrease monotonically. Hence we can upper bound its value by the sum of the first three terms, which is $(1 - \frac{c^2}{36} + \frac{c^4}{2\cdot 36^2})$. From this, we see that $1 - \exp\left(-\frac{c^2}{36}\right) \geq \frac{c^2}{36} - \frac{c^4}{2\cdot 36^2} = \frac{72c^2 - c^4}{2\cdot 36^2}$. Using this, after simplification, we see that

$$\begin{aligned}4m\cdot low \cdot \frac{1}{1-\exp\left(-\frac{c^2}{36}\right)} &\leq \frac{8\cdot 36^2 \cdot m \cdot low}{c^2\cdot(72-c^2)}\\ &\leq \frac{8\cdot 36^2\cdot m\cdot low}{71\cdot c^2} \quad \text{since } c<1\\ &\leq \frac{147\cdot m\cdot low}{c^2} \qquad (13)\end{aligned}$$

Using the bound of (13) in (12), we obtain the statement. □

We prove Theorem 12 from the main text.

THEOREM 12. *Consider the monitor M. Let $i \in \{1,\ldots,k\}$ and let $\mathcal{L}_i \subseteq \Sigma^{N\cdot m}$ be the set of observation sequences for which M outputs i. Then we have for all $j \in \{1,\ldots,k\} - \{i\}$:*

$$\begin{aligned}\mathcal{P}_i([\mathcal{L}_i\Sigma^\omega]_i) &\geq 1 - 2k\cdot\exp\left(-\frac{c^2}{18}\cdot N\right) \quad \textit{and}\\ \mathcal{P}_j([\mathcal{L}_i\Sigma^\omega]_j) &\leq 2\exp\left(-\frac{c^2}{18}\cdot N\right)\end{aligned}$$

PROOF. For $j \in \{1,\ldots,k\} - \{i\}$ define:

$$\mathcal{L}_{i,j} := \begin{cases}\{u \in \Sigma^{N\cdot m} \mid pr_i(\delta_{s_{i,0}}, u) \geq pr_j(\delta_{s_{j,0}}, u)\} & \text{if } i < j\\ \{u \in \Sigma^{N\cdot m} \mid pr_i(\delta_{s_{i,0}}, u) > pr_j(\delta_{s_{j,0}}, u)\} & \text{if } i > j\end{cases}$$

By Theorem 5 we have:

$$\begin{aligned}\mathcal{P}_i([\mathcal{L}_{i,j}\Sigma^\omega]_i) &\geq 1 - 2\cdot\exp\left(-\frac{c^2}{18}\cdot N\right) \quad \text{and}\\ \mathcal{P}_j([\mathcal{L}_{i,j}\Sigma^\omega]_j) &\leq 2\exp\left(-\frac{c^2}{18}\cdot N\right)\end{aligned} \qquad (14)$$

We have:

$$\begin{aligned}&1 - \mathcal{P}_i([\mathcal{L}_i\Sigma^\omega]_i)\\ &= 1 - \mathcal{P}_i([\cap_{j\neq i}\mathcal{L}_{i,j}\Sigma^\omega]_i) && \mathcal{L}_i = \cap_{j\neq i}\mathcal{L}_{i,j}\\ &= \mathcal{P}_i([\cup_{j\neq i}(\Sigma^{N\cdot m} - \mathcal{L}_{i,j})\Sigma^\omega]_i)\\ &\leq \sum_{j\neq i}\mathcal{P}_i([(\Sigma^{N\cdot m} - \mathcal{L}_{i,j})\Sigma^\omega]_i) && \text{union bound}\\ &= \sum_{j\neq i}\left(1 - \mathcal{P}_i([\mathcal{L}_{i,j}\Sigma^\omega]_i)\right)\\ &\leq \sum_{j\neq i}\left(2\cdot\exp\left(-\frac{c^2}{18}\cdot N\right)\right) && \text{by (14)}\\ &\leq 2k\cdot\exp\left(-\frac{c^2}{18}\cdot N\right)\end{aligned}$$

The first inequality follows. Further we have:

$$\begin{aligned}\mathcal{P}_j([\mathcal{L}_i\Sigma^\omega]_j) &\leq \mathcal{P}_j([\mathcal{L}_{i,j}\Sigma^\omega]_j) && \mathcal{L}_i \subseteq \mathcal{L}_{i,j}\\ &\leq 2\exp\left(-\frac{c^2}{18}\cdot N\right) && \text{by (14)}\end{aligned}$$

This proves the second inequality. □

B. PROOFS OF SECTION 5

We prove Proposition 14 from the main text.

PROPOSITION 14. *One can compute, in polynomial time, a test set $Test \subseteq \Sigma^*$ such that for all $\psi_1 \in Distr(S_1)$ and all $\psi_2 \in Distr(S_2)$ we have:*

$$\psi_1 \equiv \psi_2 \quad \Longleftrightarrow \quad \forall u \in Test: \ pr_1(\psi_1, u) = pr_2(\psi_2, u)$$

PROOF. For both $i = 1, 2$ and all $a \in \Sigma$ define a matrix $M_i(a) \in [0,1]^{S_i \times S_i}$ with

$$\big(M_i(a)\big)_{s,t} = \begin{cases} \phi_i(s,t) & \text{if } O_i(s) = a \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } s, t \in S_i.$$

For $i = 1, 2$ write $\eta_i \in \{1\}^{S_i}$ for the column vector all whose entries are 1. For $i = 1, 2$ and for any string $u = a_1 \dots a_k \in \Sigma^*$ define the column vector $\eta_i(u) \in [0,1]^{S_i}$ with $\eta_i(u) = M_i(a_1) \cdot \ldots \cdot M_i(a_k) \cdot \eta_i$. For all $s \in S_i$ we have, according to the definitions, the equality $\big(\eta_i(u)\big)_s = \mathcal{P}_{i,s}([u\Sigma^\omega]_i)$, which is the probability that the string u is output by H_i starting from s. For a distribution $\psi_i \in Distr(S_i)$ write $\langle \psi_i \rangle \in [0,1]^{S_i}$ for the stochastic row vector with $\langle \psi_i \rangle_s = \psi_i(s)$. According to the definitions, we have $pr_i(\psi_1, u) = \langle \psi_i \rangle \cdot \eta_i(u)$ for all $u \in \Sigma^*$. Define

$$\eta(u) := \begin{pmatrix} \eta_1(u) \\ -\eta_2(u) \end{pmatrix} \in [0,1]^{S_1 \cup S_2} \qquad \text{for all } u \in \Sigma^*.$$

Hence we have $\psi_1 \equiv \psi_2$ if and only if

$$\big(\langle \psi_1 \rangle \quad \langle \psi_2 \rangle\big) \cdot \eta(u) = 0 \qquad \text{for all } u \in \Sigma^*.$$

It follows that we have $\psi_1 \equiv \psi_2$ if and only if $\big(\langle \psi_1 \rangle \quad \langle \psi_2 \rangle\big)$ is orthogonal to the vector space, say $\mathcal{V}$, spanned by $\{\eta(u) \mid u \in \Sigma^*\}$. Define

$$M(u) := \begin{pmatrix} M_1(u) & 0 \\ 0 & M_2(u) \end{pmatrix} \in [0,1]^{(S_1 \cup S_2) \times (S_1 \cup S_2)}$$

for all $u \in \Sigma^*$. Note that $\eta(au) = M(a)\eta(u)$ holds for all $a \in \Sigma$ and all $u \in \Sigma^*$. Hence the vector space $\mathcal{V}$ can be equivalently described as the smallest vector space that contains $\eta(\varepsilon)$ (where ε denotes the empty string, i.e., all entries of $\eta(\varepsilon)$ are ± 1) and satisfies $M(a)v \in \mathcal{V}$ for all $a \in \Sigma$ and all $v \in \mathcal{V}$.

We now give a polynomial-time algorithm for computing a set $Test \subseteq \Sigma^*$. The algorithm is as follows: Initialize $Test := \{\varepsilon\}$ where ε denotes the empty string. Then, as long as there are $a \in \Sigma$ and $w \in Test$ such that $M(a)\eta(w)$ is linearly independent of $\{\eta(u) \mid u \in Test\}$, set $Test := Test \cup \{aw\}$.

Now we show that the computed set $Test$ has the properties claimed in the proposition. Since $\mathcal{V}$ is the smallest vector space that contains $\eta(\varepsilon)$ and satisfies $M(a)v \in \mathcal{V}$ for all $a \in \Sigma$ and all $v \in \mathcal{V}$, the set $U := \{\eta(u) \mid u \in Test\}$ for the computed set $Test$ is a basis for $\mathcal{V}$. Since $\mathcal{V}$ is a subspace of $\mathbb{R}^{S_1 \cup S_2}$, the dimension of $\mathcal{V}$ is at most $m = |S_1| + |S_2|$. Since U is a basis, we have $|Test| \leq m$. Since every string that the algorithm adds to $Test$ is only one letter longer than some other string already in $Test$, it follows that $|u| < |Test| \leq m$ holds for all $u \in Test$. Finally we show for all $\psi_1 \in Distr(S_1)$ and all $\psi_2 \in Distr(S_2)$:

$$\psi_1 \equiv \psi_2 \quad \Longleftrightarrow \quad \forall u \in Test: \ pr_1(\psi_1, u) = pr_2(\psi_2, u)$$

The direction "$\Longrightarrow$" is immediate. For the converse "$\Longleftarrow$", assume $pr_1(\psi_1, u) = pr_2(\psi_2, u)$ for all $u \in Test$. Then we have for all $u \in Test$:

$$\begin{aligned} 0 &= pr_1(\psi_1, u) - pr_2(\psi_2, u) \\ &= \langle \psi_1 \rangle \cdot \eta_1(u) - \langle \psi_2 \rangle \cdot \eta_2(u) \\ &= \big(\langle \psi_1 \rangle \quad \langle \psi_2 \rangle\big) \cdot \eta(u) \end{aligned}$$

Since $\{\eta(u) \mid u \in Test\} = U$ is a basis for $\mathcal{V}$, it follows that $\big(\langle \psi_1 \rangle \quad \langle \psi_2 \rangle\big)$ is orthogonal to $\mathcal{V}$. We have already argued that this implies $\psi_1 \equiv \psi_2$. This completes the proof. □

We prove Proposition 15 from the main text.

PROPOSITION 15. *Let H_1, H_2 be distinguishable HMCs. One can compute, in polynomial time, a rational number $c > 0$ such that for all reachable pairs (ψ_1, ψ_2) of distributions we have $dist(\psi_1, \psi_2) \geq c$.*

PROOF. We say a state $s_1 \in S_1$ *dominates* a distribution $\psi_1 \in Distr(S_1)$ if $\psi_1(s_1) \geq \psi_1(t_1)$ holds for all $t_1 \in S_1$. We say a pair of states (s_1, s_2) is *reachable* if there exists a reachable pair of distributions (ψ_1, ψ_2) with $\psi_i(s_i) > 0$ for both $i = 1, 2$. Note that one can compute, in polynomial time, from H_1, H_2 the set of all reachable pairs of states. For $s_1 \in S_1$ define $Unreach(s_1) := \{s_2 \in S_2 \mid (s_1, s_2)$ is not reachable.$\}$. For every $s_1 \in S_1$, consider the following linear program $\mathcal{LP}(s_1)$ over a real variable x and over real variables encoding distributions $\psi_1 \in Distr(S_1)$ and $\psi_2 \in Distr(S_2)$:

$$\begin{aligned} \text{minimize } & x \geq 0 \\ \text{subject to: } & \psi_1 \in Distr(S_1) \\ & \psi_2 \in Distr(S_2) \\ & s_1 \text{ dominates } \psi_1 \\ & \psi_2(s_2) = 0 \text{ for all } s_2 \in Unreach(s_1) \\ & -x \ \leq \ pr_1(\psi_1, u) - pr_2(\psi_2, u) \ \leq \ x \\ & \qquad \text{for all } u \in Test. \end{aligned}$$

Note that all constraints are linear (in)equalities. In particular, we have $pr_i(\psi_i, u) = \sum_{s \in S_i} \psi_i(s) \cdot \mathcal{P}_{i,s}([u\Sigma^\omega]_i)$. The probabilities $\mathcal{P}_{i,s}([u\Sigma^\omega]_i)$ can be computed in polynomial time. (Those probabilities are computed already when computing the set $Test$ according to the proof of Proposition 14: they are the probabilities in the vectors $\eta_i(u)$ defined there.)

For every $s_1 \in S_1$, let $c(s_1)$ denote the optimum solution (minimizing x) of $\mathcal{LP}(s_1)$. Define $c := \min\{c(s_1) \mid s_1 \in S_1\}$. Note that c can be computed in polynomial time. We show that c has the properties claimed by the proposition.

First we show that $dist(\psi_1, \psi_2) \geq c$ holds for all reachable pairs (ψ_1, ψ_2). Towards a contradiction suppose that there is a reachable pair (ψ_1, ψ_2) with $dist(\psi_1, \psi_2) < c$. Let $s_1 \in S_1$ be a state that dominates ψ_1. Since (ψ_1, ψ_2) is reachable, we have $\psi_2(s_2) = 0$ for all $s_2 \in Unreach(s_1)$. By the definition of $dist(\psi_1, \psi_2)$, we have

$$-dist(\psi_1, \psi_2) \ \leq \ pr_1(\psi_1, u) - pr_2(\psi_2, u) \ \leq \ dist(\psi_1, \psi_2)$$

for all $u \in Test$. It follows that $x := dist(\psi_1, \psi_2)$ along with ψ_1, ψ_2 is a feasible solution of the linear program $\mathcal{LP}(s_1)$. Since $c(s_1)$ is optimal, we have $c(s_1) \leq dist(\psi_1, \psi_2)$. By our assumption we have $dist(\psi_1, \psi_2) < c$, hence $c(s_1) < c$. But by the definition of c we have $c \leq c(s_1)$, a contradiction. We conclude that $dist(\psi_1, \psi_2) \geq c$ holds for all reachable pairs (ψ_1, ψ_2).

Finally, we show $c > 0$. Towards a contradiction suppose $c = 0$. So by definition of c there is $s_1 \in S_1$ with $c(s_1) = 0$. Thus, $\mathcal{LP}(s_1)$ has a solution with $x = 0$. That is, there exist $\psi_1 \in \mathit{Distr}(S_1)$ and $\psi_2 \in \mathit{Distr}(S_2)$ such that s_1 dominates ψ_1, and

$$\psi_2(s_2) = 0 \text{ holds for all } s_2 \in \mathit{Unreach}(s_1), \qquad (15)$$

and $pr_1(\psi_1, u) = pr_2(\psi_2, u)$ holds for all $u \in \mathit{Test}$. By Proposition 14, the last fact implies

$$\psi_1 \equiv \psi_2 \,. \qquad (16)$$

Since s_1 dominates ψ_1, we have

$$\psi_1(s_1) > 0 \,. \qquad (17)$$

It follows directly from [7, Theorem 21] that (15)–(17) together imply that we have $d(H_1, H_2) < 1$ for the the total variation distance d defined in the beginning of Section 3. But then Proposition 1 implies that H_1, H_2 are not distinguishable, which is a contradiction. Hence $c > 0$ must hold. This concludes the proof. □

We prove Proposition 17 from the main text.

PROPOSITION 17. *Let H_1, H_2 be two distinguishable HMCs. One can compute, in exponential time:*

$$c := \min_{\text{reachable } (S_1', S_2') \in 2^{S_1} \times 2^{S_2}} \; \min_{\psi_1 \in \mathit{Distr}(S_1')} \; \min_{\psi_2 \in \mathit{Distr}(S_2')} \; \max_{U \subseteq \Sigma^m} \big(pr_1(\psi_1, U) - pr_2(\psi_2, U) \big)$$

PROOF. The reachable pairs $(S_1', S_2') \in 2^{S_1} \times 2^{S_2}$ can be computed in exponential time. So it suffices to show that one can compute, for a fixed reachable pair $(S_1', S_2') \in 2^{S_1} \times 2^{S_2}$, the value

$$c_{S_1', S_2'} := \min_{\psi_1 \in \mathit{Distr}(S_1')} \; \min_{\psi_2 \in \mathit{Distr}(S_2')} \; \max_{U \subseteq \Sigma^m} \big(pr_1(\psi_1, U) - pr_2(\psi_2, U) \big)$$

in exponential time. Consider the following linear program, similar to the one from the proof of Proposition 15, with variables x_u for $u \in \Sigma^m$ and variables encoding distributions ψ_1, ψ_2:

$$\begin{aligned} \text{minimize } & \sum_{u \in \Sigma^m} x_u \\ \text{subject to: } & \psi_1 \in \mathit{Distr}(S_1) \\ & \psi_2 \in \mathit{Distr}(S_2) \\ & 0 \le x_u && \text{for all } u \in \Sigma^m \\ & pr_1(\psi_1, u) - pr_2(\psi_2, u) \;\le\; x_u && \text{for all } u \in \Sigma^m \end{aligned}$$

This linear program has exponential size. We show that its optimal solution is $c_{S_1', S_2'}$.

First we show that it has a feasible solution whose value is $c_{S_1', S_2'}$. Let ψ_1, ψ_2 be the distributions that attain the minimum from the definition of $c_{S_1', S_2'}$. Let U be a set that attains the maximum from the definition of $c_{S_1', S_2'}$. We can take $U = \{u \in \Sigma^m \mid pr_1(\psi_1, u) \ge pr_2(\psi_2, u)\}$. Let $x_u = pr_1(\psi_1, u) - pr_2(\psi_2, u)$ for all $u \in U$, and let $x_u = 0$ for all $u \in \Sigma^m - U$. Then the solution with those x_u and with ψ_1, ψ_2 is feasible. Moreover, its value is:

$$\begin{aligned} \sum_{u \in \Sigma^m} x_u &= \sum_{u \in U} x_u \\ &= \sum_{u \in U} \big(pr_1(\psi_1, u) - pr_2(\psi_2, u) \big) \\ &= pr_1(\psi_1, U) - pr_2(\psi_2, U) \\ &= c_{S_1', S_2'} \end{aligned}$$

For the converse, we show that $c_{S_1', S_2'}$ is a lower bound to the value of any feasible solution. Let $(x_u)_{u \in \Sigma^m}$ along with ψ_1, ψ_2 denote a feasible solution. Let U be a set that attains the maximum in $\max_{U \subseteq \Sigma^m} \big(pr_1(\psi_1, U) - pr_2(\psi_2, U) \big)$. We can take $U = \{u \in \Sigma^m \mid pr_1(\psi_1, u) \ge pr_2(\psi_2, u)\}$. Hence we have:

$$\begin{aligned} & \sum_{u \in \Sigma^m} x_u \\ \ge\; & \sum_{u \in U} x_u \\ & \qquad (x_u \ge 0 \text{ from the linear program}) \\ \ge\; & \sum_{u \in U} \big(pr_1(\psi_1, u) - pr_2(\psi_2, u) \big) \\ & \qquad (x \ge pr_1(\psi_1, u) - pr_2(\psi_2, u) \text{ from the lin. program}) \\ =\; & pr_1(\psi_1, U) - pr_2(\psi_2, U) \\ =\; & \max_{U \subseteq \Sigma^m} \big(pr_1(\psi_1, U) - pr_2(\psi_2, U) \big) \\ \ge\; & c_{S_1', S_2'} \\ & \qquad \Big(\text{definition of } c_{S_1', S_2'} \Big) \end{aligned}$$

We conclude that $c_{S_1', S_2'}$ is an optimal solution of the linear program. □

C. PROOFS OF SECTION 6

We prove Proposition 18 from the main text.

PROPOSITION 18. *Let H be a cHMC with $\mathcal{P}(\mathit{Bad}), \mathcal{P}(\mathit{Good}) > 0$. Then one can compute, in polynomial time, HMCs H_1, H_2 such that for all measurable events $E \subseteq S^\omega$ we have*

$$\mathcal{P}_1(E) = \mathcal{P}(E \mid \mathit{Bad}) \quad \textit{and} \quad \mathcal{P}_2(E) = \mathcal{P}(E \mid \mathit{Good}) \,.$$

PROOF. By symmetry, it suffices to provide the construction for H_1. Let $H = (G, O, s_0, \mathit{Class})$ be the given cHMC with $G = (S, R, \phi)$ a Markov chain. Define

$$S_1 := \{s \in S \mid \mathcal{P}_s(\mathit{Bad}_s) > 0\} \,.$$

Note that $s_0 \in S_1$. Define $G_1 := (S_1, R_1, \phi_1)$ with $R_1 := R \cap (S_1 \times S_1)$ and

$$\phi_1(s, t) := \frac{\phi(s, t) \cdot \mathcal{P}_t(\mathit{Bad}_t)}{\mathcal{P}_s(\mathit{Bad}_s)} \qquad \text{for all } (s, t) \in R_1 .$$

Finally, take $H_1 := (G_1, O_1, s_0)$ where O_1 equals O restricted to S_1.

We show that the measures $\mathcal{P}_1(\cdot)$ and $\mathcal{P}(\cdot \mid \mathit{Bad})$ are equal. By definition, it suffices to show that they are equal on the cylinder sets $\{s_0 r\} S_1^\omega$ for all $r \in S_1^*$. We show by induction on the length of r that

$$\mathcal{P}_{1,s}(\{sr\} S_1^\omega) \cdot \mathcal{P}_s(\mathit{Bad}_s) = \mathcal{P}_s(\{sr\} S_1^\omega \cap \mathit{Bad}_s) \quad \forall\, s \in S_1 .$$

For the induction base, let r be empty. Then the claim follows from $\mathcal{P}_{1,s}(\{s\}S_1^\omega) = 1$ and $Bad_s \subseteq \{s\}S_1^\omega$. For the induction step, let $t \in S_1$ and $r \in S_1^*$. We want to show:

$$\mathcal{P}_{1,s}(\{str\}S_1^\omega) \cdot \mathcal{P}_s(Bad_s) = \mathcal{P}_s(\{str\}S_1^\omega \cap Bad_s) \qquad (18)$$

If $(s,t) \notin R_1$ then both sides of (18) are zero. So let $(s,t) \in R_1$. Then we have:

$$\begin{aligned}
&\mathcal{P}_{1,s}(\{str\}S_1^\omega) \cdot \mathcal{P}_s(Bad_s) \\
&= \phi_1(s,t) \cdot \mathcal{P}_{1,t}(\{tr\}S_1^\omega) \cdot \mathcal{P}_s(Bad_s) \\
&= \frac{\phi(s,t) \cdot \mathcal{P}_t(Bad_t)}{\mathcal{P}_s(Bad_s)} \cdot \mathcal{P}_{1,t}(\{tr\}S_1^\omega) \cdot \mathcal{P}_s(Bad_s) \\
&= \phi(s,t) \cdot \mathcal{P}_t(\{tr\}S_1^\omega \cap Bad_t) \qquad \text{by the ind. hyp.} \\
&= \mathcal{P}_s(\{str\}S_1^\omega \cap Bad_s)
\end{aligned}$$

This shows (18) and hence the proposition. □