Equivalence of Hidden Markov Models with Continuous Observations

Oscar Darwin
Department of Computer Science, Oxford University, United Kingdom

Stefan Kiefer
Department of Computer Science, Oxford University, United Kingdom

Abstract

We consider Hidden Markov Models that emit sequences of observations that are drawn from continuous distributions. For example, such a model may emit a sequence of numbers, each of which is drawn from a uniform distribution, but the support of the uniform distribution depends on the state of the Hidden Markov Model. Such models generalise the more common version where each observation is drawn from a finite alphabet. We prove that one can determine in polynomial time whether two Hidden Markov Models with continuous observations are equivalent.

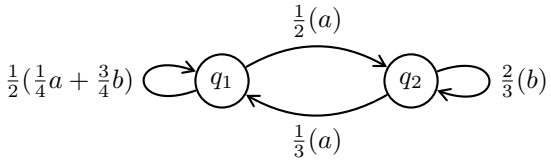
2012 ACM Subject Classification Theory of computation → Random walks and Markov chains; Mathematics of computing → Stochastic processes; Theory of computation → Logic and verification

Keywords and phrases Markov chains, equivalence, probabilistic systems, verification

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

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. For any two states q, q', whenever the state changes from q to q', the HMM samples and then emits a random observation according to a probability distribution $D(q, q')$. For example, consider the following diagram visualising a HMM:



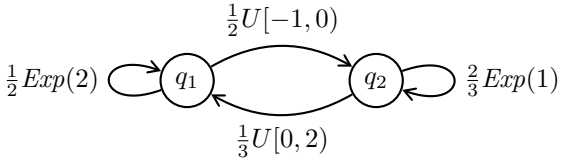
In state q_1, the successor state is q_1 or q_2, with probability $\frac{1}{2}$ each. Upon transitioning from q_1 to itself, observation a is drawn with probability $\frac{1}{4}$ and observation b is drawn with probability $\frac{3}{4}$; upon transitioning from q_1 to q_2, observation a is drawn surely.[1]

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

In the example above the observations are drawn from a finite observation alphabet $\Sigma = \{a, b\}$. Indeed, in the literature HMMs most commonly have a finite observation alphabet. In this paper we lift this restriction and consider *continuous-observation* HMMs, by which we mean HMMs as described above, but with continuous observation set Σ. For example,

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

instead of the distributions on $\{a, b\}$ in the picture above (written there as $(\frac{1}{4}a + \frac{3}{4}b)$, (a), (b), respectively), we may have distributions on the real numbers. For example in the following diagram, where $U[a, b)$ denotes the uniform distribution on $[a, b)$ and $Exp(\lambda)$ denotes the exponential distribution with parameter λ:



HMMs, both with finite and infinite observation sets, are widely employed in fields such as speech recognition (see [22] for a tutorial), gesture recognition [7], signal processing [11], and climate modeling [1]. HMMs are heavily used in computational biology [14], more specifically in DNA modeling [9] and biological sequence analysis [13], including protein structure prediction [18] and gene finding [2]. In computer-aided verification, HMMs are the most fundamental model for probabilistic systems; model-checking tools such as Prism [19] and Storm [12] are based on analyzing HMMs efficiently.

One of the most fundamental questions about HMMs is whether two HMMs with initial state distributions are *(trace) equivalent*, i.e., generate the same distribution on infinite observation sequences. For finite observation alphabets this problem is very well studied and can be solved in polynomial time using algorithms that are based on linear algebra [23, 21, 24, 10]. Checking trace equivalence is used in the verification of obliviousness and anonymity, properties that are hard to formalize in temporal logics, see, e.g., [3, 17, 5].

Although the generalisation to continuous observations (such as passed time, consumed energy, sensor readings) is natural, there has been little work on the algorithmics of such HMMs. One exception is *continuous-time* Markov chains (CTMCs) [4, 8] which are similar to HMMs described above, but with two kinds of observations: on the one hand they emit observations from a finite alphabet, but on the other hand they also emit the *time* spent in each state. Typically, each state-to-state transition is labelled with a parameter λ; for each transition its time of "firing" is drawn from an exponential distribution with parameter λ; the transition with the smallest firing time "wins" and causes the corresponding change of state. CTMCs have attractive properties: they are in a sense memoryless, and for many analyses, including model checking, an equivalent discrete-time model can be calculated using an efficient and numerically stable process called *uniformization* [15].

In [16] a stochastic model more general than ours was introduced, allowing not only for uncountable sets of observations (called *labels* there), but also for infinite sets of states and actions. The paper [16] focuses on bisimulation; trace equivalence is not considered. It emphasizes nondeterminism, a feature we do not consider here.

To the best of the authors' knowledge, this paper is the first to study equivalence of HMMs with continuous observations. As continuous functions are part of the input, an equivalence checking algorithm, if it exists (which is not a priori clear), needs to be *symbolic*, i.e., needs to perform computations on functions. Our contributions are as follows:

1. We show in Section 3 that certain aspects of the linear-algebra based approach for checking equivalence of finite-observation HMMs carry over to the continuous case naturally. In particular, equivalence reduces to orthogonality in a certain vector space of state-indexed real vectors, see Proposition 7.
2. However, we show in Section 4 that in the continuous case there can be *additional* linear dependencies between the observation density functions (which is impossible in the finite

case, where the different observations can be assumed linearly independent). This renders a simple-minded reduction to the finite case incorrect. Therefore, an equivalence checking algorithm needs to consider the interplay with the vector space from item 1.

3. For the required computations on the observation density functions we introduce in Section 5 *linearly decomposable profile languages*, which are languages (i.e., sets of finite words) whose elements encode density functions on which basis computations can be performed efficiently. In Section 5.1 we provide an extensive example of such a language, encoding (linear combinations of) Gaussian, exponential, and piecewise polynomial density functions. The proof that this language has the required properties is non-trivial itself and requires *alternant matrices* and comparisons of the tails of various density functions.
4. In Section 6 we finally show that HMMs whose observation densities are given in terms of linearly decomposable profile languages can be checked for equivalence in *polynomial time*, by a reduction to the finite-observation case. We also indicate, in Example 23, how our result can be used to check for susceptibility of certain timing attacks.

2 Preliminaries

We write $\mathbb{N}$ for the set of positive integers, $\mathbb{Q}$ for the set of rationals and $\mathbb{Q}_+$ for the set of positive rationals. For $d \in \mathbb{N}$ and a finite set Q we use the notation $|Q|$ for the number of elements in Q, $[d] = \{1, \ldots, d\}$ and $[Q] = \{1, \ldots, |Q|\}$. Vectors $\mu \in \mathbb{R}^N$ are viewed as row vectors and we write $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^N$. Superscript T denotes transpose; e.g., $\mathbb{1}^T$ is a column vector of ones. A matrix $M \in \mathbb{R}^{N \times N}$ is *stochastic* if M is non-negative and $\sum_{j=1}^{N} M_{i,j} = 1$ for all $i \in [N]$. For a domain Σ and subset $E \subseteq \Sigma$ the *characteristic* function $\chi_E : \Sigma \to \{0, 1\}$ is defined as $\chi_E(x) = 1$ if $x \in E$ and $\chi_E(x) = 0$ otherwise.

Throughout this paper, we use Σ to denote a set of *observations*. We assume Σ is a topological space and $(\Sigma, \mathcal{G}, \lambda)$ is a measure space where all the open subsets of Σ are contained within $\mathcal{G}$ and have non-zero measure. Indeed $\mathbb{R}$ and the usual Lebesgue measure space on $\mathbb{R}$ satisfy these assumptions. The set Σ^n is the set of words over Σ of length n and $\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$.

A matrix valued function $\Psi : \Sigma \to [0, \infty)^{N \times N}$ can be integrated element-wise. We write $\int_E \Psi \, d\lambda$ for the matrix with entries $\left(\int_E \Psi \, d\lambda\right)_{i,j} = \int_E \Psi_{i,j} \, d\lambda$, where $\Psi_{i,j} : \Sigma \to [0, \infty)$ is defined by $\Psi_{i,j}(x) = \big(\Psi(x)\big)_{i,j}$ for all $x \in \Sigma$.

A function $f : \Sigma \to \mathbb{R}^m$ is *piecewise continuous* if there is an open set $C \subseteq \Sigma$, called a *set of continuity*, such that f is continuous on C and for every point $x \in \Sigma \setminus C$ there is some sequence of points $x_n \in C$ such that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} f(x_n) = f(x)$. For a non-negative function $f : \Sigma \to [0, \infty)$ we use the notation $\operatorname{supp} f = \{x \in \Sigma \mid f(x) > 0\}$.

▶ **Definition 1.** *A* Hidden Markov Model *(HMM) is a triple (Q, Σ, Ψ) where Q is a finite set of states, Σ is a set of observations, and the* observation density matrix $\Psi : \Sigma \to [0, \infty)^{|Q| \times |Q|}$ *specifies the transitions such that $\int_\Sigma \Psi \, d\lambda$ is a stochastic matrix.*

▶ **Example 2.** The second HMM from the introduction is the triple $(\{q_1, q_2\}, \mathbb{R}, \Psi)$ with

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

We assume that Ψ is piecewise continuous and extend Ψ to the mapping $\Psi : \Sigma^* \to [0, \infty)^{|Q| \times |Q|}$ with $\Psi(x_1 \cdots x_n) = \Psi(x_1) \times \cdots \times \Psi(x_n)$ for $x_1, \ldots, x_n \in \Sigma$. If C is the set of continuity for $\Psi : \Sigma \to [0, \infty)^{|Q| \times |Q|}$, then for fixed $n \in \mathbb{N}$ the restriction $\Psi : \Sigma^n \to [0, \infty)^{|Q| \times |Q|}$ is piecewise continuous with set of continuity C^n. We say that $A \subseteq \Sigma^n$ is a *cylinder set*

if $A = A_1 \times \cdots \times A_n$ and $A_i \in \mathcal{G}$ for $i \in [n]$. For every n there is an induced measure space $(\Sigma^n, \mathcal{G}^n, \lambda^n)$ where $\mathcal{G}^n$ is the smallest σ-algebra containing all cylinder sets in Σ^n and $\lambda^n(A_1 \times \cdots \times A_n) = \prod_{i=1}^n \lambda(A_i)$ for any cylinder set $A_1 \times \cdots \times A_n$. Let $A \subseteq \Sigma^n$ and write $A\Sigma^\omega$ for the set of infinite words over Σ where the first n observations fall in the set A. Given a HMM (Q, Σ, Ψ) and initial distribution π on Q viewed as vector $\pi \in \mathbb{R}^{|Q|}$, there is an induced probability space $(\Sigma^\omega, \mathcal{G}^*, \mathbb{P}_\pi)$ where Σ^ω is the set of infinite words over Σ, and $\mathcal{G}^*$ is the smallest σ-algebra containing (for all $n \in \mathbb{N}$) all sets $A\Sigma^\omega$ where $A \subseteq \Sigma^n$ is a cylinder set and $\mathbb{P}_\pi$ is the unique probability measure such that $\mathbb{P}_\pi(A\Sigma^\omega) = \pi \int_A \Psi \, d\lambda^n \mathbb{1}^T$ for any cylinder set $A \subseteq \Sigma^n$.

▶ **Definition 3.** *For two distributions π_1 and π_2 and a HMM $C = (Q, \Sigma, \Psi)$, we say that π_1 and π_2 are* equivalent*, written $\pi_1 \equiv_C \pi_2$, if $\mathbb{P}_{\pi_1}(A) = \mathbb{P}_{\pi_2}(A)$ holds for all measurable subsets $A \subseteq \Sigma^\omega$.*

One could define equivalence of two pairs (C_1, π_1) and (C_2, π_2) where $C_i = (Q_i, \Sigma, \Psi_i)$ are HMMs and π_i are initial distributions for $i = 1, 2$. We do not need that though, as we can define, in a natural way, a single HMM over the disjoint union of Q_1 and Q_2 and consider instead equivalence of π_1 and π_2 (where π_1, π_2 are appropriately padded with zeros).

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

▶ **Example 4.** The observation density matrix Ψ from Example 2 has a functional decomposition

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

◀

▶ **Lemma 5.** *Let (Q, Σ, Ψ) be a HMM. If Ψ has functional decomposition $\Psi = \sum_{k=1}^d f_k P_k$ then $\sum_{k=1}^d P_k$ is stochastic.*

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

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

◀

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

Encoding For computational purposes we assume that rational numbers are represented as ratios of integers in binary. The initial distribution of a HMM with state set Q is given as a vector $\pi \in \mathbb{Q}^{|Q|}$. We also need to encode continuous functions, in particular, density functions such as Gaussian, exponential or piecewise-polynomial functions. A *profile* is a finite word (i.e., string) that describes a continuous function. It may consist of (an encoding of) a function type and its parameters. For example, the profile $(\mathcal{N}, \mu, \sigma)$ may denote a Gaussian (also called normal) distribution with mean $\mu \in \mathbb{Q}$ and standard deviation $\sigma \in \mathbb{Q}_+$. A profile may also consist of a description of a rational linear combination of such building blocks. For any profile γ we write $[\![\gamma]\!] : \Sigma \to [0, \infty)$ for the function it encodes. For example,

a profile $\gamma = (\mathcal{N}, \mu, \sigma)$ with $\mu \in \mathbb{Q}$, $\sigma \in \mathbb{Q}_+$ may encode the function $[\![\gamma]\!] : \mathbb{R} \to [0, \infty)$ given as $[\![\gamma]\!](x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$. Without restricting ourselves to any particular encoding, we assume that Γ is a *profile language*, i.e., a finitely presented but usually infinite set of valid profiles. For any $\Gamma_0 \subseteq \Gamma$ we write $[\![\Gamma_0]\!] = \{[\![\gamma]\!] \mid \gamma \in \Gamma_0\}$.

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

▶ **Example 6.** For a suitable profile language Γ, the HMM from Example 2 may be over Γ, with the observation density matrix given as

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

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

3 Equivalence as Orthogonality

For finite-observation HMMs it is well known [23, 21, 24, 10] that two initial distributions given as vectors $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ are equivalent if and only if $\pi_1 - \pi_2$ is orthogonal (written as $\perp$) to a certain vector space. Indeed, this property holds more generally:

▶ **Proposition 7.** *Consider a HMM (Q, Σ, Ψ). For any $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ we have*

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

In the finite-observation case, Proposition 7 leads to an efficient algorithm for deciding equivalence: it suffices to compute a basis for $\mathcal{V} = \operatorname{span}\{\Psi(w)\mathbb{1}^T \mid w \in \Sigma^*\}$. This can be done using a fixed-point algorithm that computes a sequence of (bases of) increasing subspaces of $\mathcal{V}$: start with $\mathcal{B} = \{\mathbb{1}^T\}$, and as long as there is $a \in \Sigma$ and $v \in \mathcal{B}$ such that $\Psi(a)v \notin \operatorname{span} \mathcal{B}$, add $\Psi(a)v$ to $\mathcal{B}$. Since $\dim \mathcal{V} \leq |Q|$, this algorithm terminates after at most $|Q|$ iterations, and returns $\mathcal{B}$ such that $\operatorname{span} \mathcal{B} = \mathcal{V}$. It is then easy to check whether $\pi_1 - \pi_2 \perp \mathcal{V}$. It follows:

▶ **Proposition 8.** *Given a HMM (Q, Σ, Ψ) with finite Σ and initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^{|Q|}$, it is decidable in polynomial time whether $\pi_1 \equiv \pi_2$.*

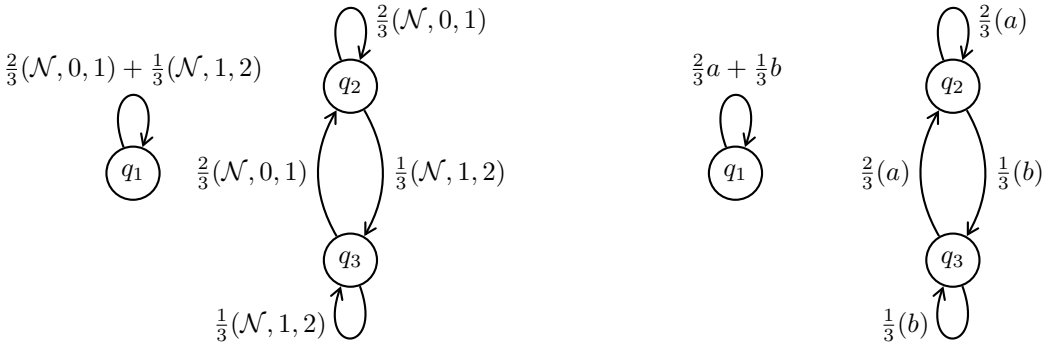
This is not an effective algorithm when Σ is infinite.

4 Labelling Reductions

Our goal is to reduce in polynomial time the equivalence problem in continuous-observation HMMs to the equivalence problem in finite-observation HMMs. Since the latter is decidable in polynomial time by Proposition 8, a polynomial time algorithm for deciding equivalence in continuous-observation HMMs follows.

Towards this objective, consider a reduction where each continuous density function is given a label and these labels form the observation alphabet of a finite-observation HMM. For example consider the chain on the left in the diagram below. This disconnected HMM

emits letters from two distinct normal distributions with profiles $(\mathcal{N}, 0, 1)$ and $(\mathcal{N}, 1, 2)$. Assigning each distribution letters a, b respectively yields the HMM given on the right. Since in the right chain states q_1 and q_2 are equivalent so too are the same labelled states in the continuous chain.



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

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

Since Ψ has functional decomposition $\Psi = \sum_{k=1}^{K} [\![\beta_k]\!] \hat{M}(a_k)$, it follows by Lemma 5 that $\sum_{k=1}^{K} \hat{M}(a_k)$ is stochastic and the labelling reduction is a well defined HMM which may be computed in polynomial time. As discussed in the previous example, equivalence in the labelling reduction implies equivalence in the original chain:

▶ **Proposition 9.** *Let $C = (Q, \Sigma, \Psi)$ be a HMM with labelling reduction $L = (Q, \hat{\Sigma}, \hat{M})$. Then for any initial distributions π_1 and π_2*

$$\pi_1 \equiv_L \pi_2 \implies \pi_1 \equiv_C \pi_2.$$

For the proof of Proposition 9 we use the following lemma which will be re-used in Section 6.

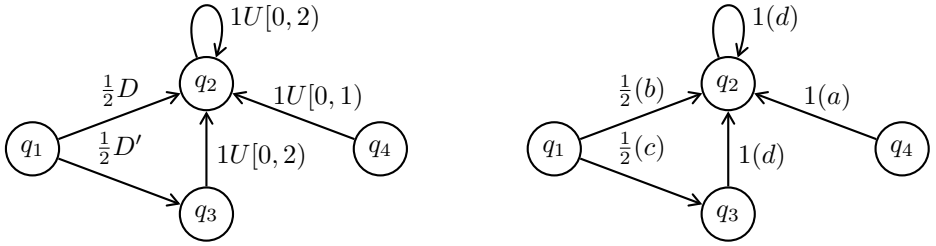
▶ **Lemma 10.** *Let $C_1 = (Q, \Sigma_1, \Psi_1)$ and $C_2 = (Q, \Sigma_2, \Psi_2)$ be two HMMs with the same state space Q. Suppose that* $\operatorname{span}\{\Psi_1(x) \mid x \in \Sigma_1\} \subseteq \operatorname{span}\{\Psi_2(x) \mid x \in \Sigma_2\}$. *Then, for any two initial distributions π_1 and π_2,*

$$\pi_1 \equiv_{C_2} \pi_2 \implies \pi_1 \equiv_{C_1} \pi_2.$$

Proof of Proposition 9. Ψ has a functional decomposition $\Psi = \sum_{k=1}^{K} [\![\beta_k]\!] \hat{M}(a_k)$. Thus, $\operatorname{span}\{\Psi(x) \mid x \in \Sigma\} \subseteq \operatorname{span}\{\hat{M}(a_k) \mid a_k \in \hat{\Sigma}\}$ and the statement follows by Lemma 10. ◀

▶ **Example 11.** Consider the HMMs in the diagram below. The HMM on the left is a continuous-observation chain where D and D' are distributions on $[0, 1]$ with probability density functions $2x\chi_{[0,1)}(x)$ and $2(1-x)\chi_{[0,1)}(x)$ respectively, and $U[a, b)$ is the uniform distribution on $[a, b)$. The HMM on the right is the corresponding labelling reduction.

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



5 Linearly Decomposable Profile Languages

Example 11 shows that the linear combination of two continuous distributions can "imitate" a single distribution. Therefore we consider the transition densities as part of a vector space of functions. In the usual way $\mathcal{L}_1(\Sigma, \lambda)$ is the quotient vector space where functions that differ only on a λ-null set are identified. In particular, when $\Sigma \subseteq \mathbb{R}$ and λ is the Lebesgue measure λ_{Leb}, the functions $\chi_{[a,b)}$ and $\chi_{(a,b]}$ are considered the same.

Let Γ be a profile language with $[\![\Gamma]\!] \subseteq \mathcal{L}_1(\Sigma, \lambda)$. We say that Γ is *linearly decomposable* if for every finite set $\{\gamma_1, \ldots, \gamma_n\} = \Gamma_0 \subseteq \Gamma$ one can compute in polynomial time profiles $\beta_1, \ldots, \beta_m \in \Gamma_0$ such that $\{[\![\beta_1]\!], \ldots, [\![\beta_m]\!]\}$ is a basis for span $\{[\![\gamma_1]\!], \ldots, [\![\gamma_n]\!]\}$ (hence $m \leq n$), and further a set of coefficients $b_{i,j} \in \mathbb{Q}$ for $i \in [n], j \in [m]$ such that

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

The following theorem is the main result of this paper:

▶ **Theorem 12.** *Given a HMM (Q, Σ, Ψ) over a linearly decomposable profile language, and initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^{|Q|}$, it is decidable in polynomial time (in the size of the encoding) whether $\pi_1 \equiv \pi_2$.*

We prove Theorem 12 in Section 6. To make the notion of linearly decomposable profile languages more concrete, we give a concrete example in the following subsection.

5.1 Example: Gaussian, Exponential, and Piecewise Polynomial Functions

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

We call a function of the form $x \mapsto x^k \chi_I(x)$ where $k \in \mathbb{N} \cup \{0\}$ and $I \subset \mathbb{R}$ is an interval an *interval-domain monomial*. To avoid clutter, we often denote interval-domain monomials only by $x^k \chi_I$. Recall that $\mathcal{L}_1(\mathbb{R}, \lambda_{Leb})$ is a quotient space, so half open intervals $I = [a, b)$ are sufficient. Any piecewise polynomial is a linear combination of interval-domain monomials.

Let M be a set of profiles encoding interval-domain monomials $x^k \chi_{[a,b)}$ in terms of $k \in \mathbb{N} \cup \{0\}$ and $a, b \in \mathbb{Q}$. Gaussian and exponential density functions can be fully described using their parameters, which we assume to be rational. We write G and E for corresponding sets of profiles, respectively. Finally, we fix a profile language $\Gamma_{GEM} \supset G \cup E \cup M$ obtained by closing $G \cup E \cup M$ under linear combinations. That is, for any $\gamma_1, \ldots, \gamma_k \in \Gamma_{GEM}$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{Q}$, there exists a profile $\gamma \in \Gamma_{GEM}$ such that $[\![\gamma]\!] = \lambda_1 [\![\gamma_1]\!] + \cdots + \lambda_k [\![\gamma_k]\!]$. This closure can be achieved using a specific constructor, say $\mathcal{S}$, for linear combinations, so that $\gamma = \mathcal{S}(\lambda_1, \gamma_1, \ldots, \lambda_k, \gamma_k)$.

▶ **Example 13.** The HMM $(Q, \mathbb{R}, \Psi)$ from Example 11 is over Γ_{GEM}: the observation density matrix Ψ can be encoded as a matrix of coefficient-profile pairs

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

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

▶ **Lemma 14.** *Let H be a set of disjoint half open intervals. Suppose that $m_1, \ldots, m_I$ are distinct interval-domain monomials such that* supp $m_i \in H$ *for all $i \in [I]$. In addition, let $g_1, \ldots, g_J$ and $e_1, \ldots, e_K$ be distinct Gaussian and exponential density functions, respectively. Then, the set $\{m_1, \ldots, m_I, g_1, \ldots, g_J, e_1, \ldots, e_K\}$ is linearly independent.*

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

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

is called the alternant matrix for $f_1, \ldots, f_n$ and *input points* $x_1, \ldots, x_n$.

▶ **Lemma 15.** *Suppose $f_1, \ldots, f_n \in \mathcal{L}_1(\Sigma, \lambda)$. Then, the f_i are linearly dependent if and only if all alternant matrices for the f_i are singular.*

Sketch proof of Lemma 14. Under the assumption that a linear combination exists almost surely equal to 0, by examining the limit at $+\infty$ we show that the exponential and Gaussian coefficients are zero. Then, by constructing an appropriate alternant matrix with full rank we invoke Lemma 15 which means the remaining interval-domain monomials are linearly independent and thus must also have zero coefficients. ◀

▶ **Proposition 16.** *The profile language Γ_{GEM} is linearly decomposable.*

Thus we obtain the following corollary of Theorem 12:

▶ **Corollary 17.** *Given a HMM (Q, Σ, Ψ) over Γ_{GEM}, and initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^{|Q|}$, it is decidable in polynomial time whether $\pi_1 \equiv \pi_2$.*

6 Proof of Theorem 12

Suppose that Ψ has a functional decomposition $\sum_{k=1}^{d} f_k P_k$ such that the set $\{f_1, \ldots, f_d\}$ is linearly independent. Then, $\sum_{k=1}^{d} f_k P_k$ is called an *independent functional decomposition*. The efficient computation of an independent functional decomposition is the key ingredient for the proof of Theorem 12. We start with the following lemma.

▶ **Lemma 18.** *Suppose $\Psi : \Sigma \to [0, \infty)^{|Q| \times |Q|}$ has an independent functional decomposition $\Psi = \sum_{k=1}^{d} f_k P_k$. Then,* span $\{\Psi(x) \mid x \in \Sigma\}$ = span $\{P_k \mid k \in [d]\}$.

Proof. Since $\Psi(x) = \sum_{k=1}^{d} f_k(x) P_k$, we have $\operatorname{span}\{\Psi(x) \mid x \in \Sigma\} \subseteq \operatorname{span}\{P_k \mid k \in [d]\}$. For the reverse inclusion, since the f_i are linearly independent, by Lemma 15 there exists an alternant matrix M with full rank for $f_1, \ldots, f_d$ with input points $x_1, \ldots, x_d$. Hence, for each of the standard basis vectors $e_k \in \{0,1\}^d$, $k \in [d]$, there exists $v_k = (v_{1,k}, \ldots, v_{d,k}) \in \mathbb{R}^d$ such that $v_k M = e_k$. Writing $\delta_{j,k}$ for the Kronecker delta function it follows that

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

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

The proof of the following proposition re-uses Lemma 10 from Section 4.

▶ **Proposition 19.** *Suppose that HMM $C = (Q, \Sigma, \Psi)$ has independent functional decomposition $\Psi = \sum_{k=1}^{d} f_k P_k$ and each P_k is non-negative for all $k \in [d]$. Define a set $\overline{\Sigma} = \{a_1, \ldots, a_d\}$ of fresh observations and the observation density matrix M with $M(a_k) = P_k$ for all $k \in [d]$. Then $F = (Q, \overline{\Sigma}, M)$ is a finite-observation HMM and for any initial distributions π_1, π_2*

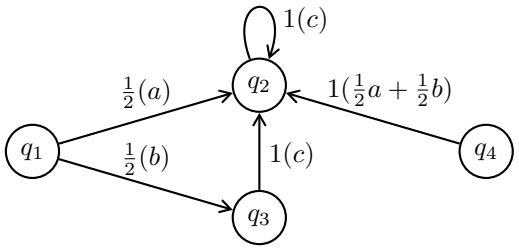
$$\pi_1 \equiv_C \pi_2 \iff \pi_1 \equiv_F \pi_2.$$

Proof. It follows by Lemma 5 that $\sum_{k=1}^{d} P_k$ is stochastic. Thus F defines a HMM. By Lemma 18, $\operatorname{span}\{\Psi(x)\mathbb{1}^T \mid x \in \Sigma\} = \operatorname{span}\{M(a)\mathbb{1}^T \mid a \in \overline{\Sigma}\}$ which combined with Lemma 10 gives the result. ◀

▶ **Example 20.** We use the HMM C discussed in Examples 11 and 13 to illustrate the construction of Proposition 19. The basis $\{2x\chi_{[0,1)}, 2(1-x)\chi_{[0,1)}, \frac{1}{2}\chi_{[0,2)}\}$ leads to the independent functional decomposition

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

Therefore, Proposition 19 implies that two initial distributions $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ are equivalent in C if and only if they are equivalent in the following HMM:



Here, states q_1 and q_4 are equivalent. Hence, they are also equivalent in C. ◀

If an observation density matrix has an entry with pdf $2e^{-x} - 2e^{-2x}$ (which is encodable in Γ_{GEM} due to its convex closure property), the independent functional decomposition generated by the algorithm described in the proof of Proposition 16 in the appendix has matrices which are not all non-negative. Therefore, Proposition 19 cannot be applied directly. However, given an independent functional decomposition $\Psi = \sum_{k=1}^{d} f_k P_k$ and noting that

$\sum_{k=1}^{d} P_k$ is stochastic by Lemma 5, the following proposition shows that there is a small $\theta > 0$ such that $P - \theta P_k$ is non-negative for all $k \in [d]$. Furthermore, span $\{P_k \mid k \in [d]\} =$ span $\{P - \theta P_k \mid k \in [d]\}$. These two facts lead us to construct a finite-observation HMM using the scaled transition matrices $\frac{1}{d-\theta}(P - \theta P_k)$.

▶ **Proposition 21.** *Let $C = (Q, \Sigma, \Psi)$ be a HMM with independent functional decomposition $\Psi = \sum_{k=1}^{d} f_k P_k$. Let $P = \sum_{k=1}^{d} P_k$ and*

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

Define an alphabet $\tilde{\Sigma} = \{a_1, \ldots, a_d\}$ of fresh observations and the HMM $F = (Q, \tilde{\Sigma}, M)$ with $M(a_k) = \frac{1}{d-\theta}(P - \theta P_k)$. Then, for any initial distributions μ_1, μ_2

$$\mu_1 \equiv_F \mu_2 \iff \mu_1 \equiv_C \mu_2.$$

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

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

and by Lemma 5, P is stochastic. In addition we must show that $M(a_k)$ is non-negative for each $k \in [d]$. Since $\theta \leq \frac{1}{2}$, it is enough to show that $P - \theta P_k$ is non-negative for each $k \in [d]$. Suppose that $(P)_{i,j} = 0$. Then, $\int_\Sigma \Psi_{i,j}\, d\lambda = (P)_{i,j} = 0$, which implies that $\Psi_{i,j} = 0$ since Ψ is piecewise continuous. Thus, $\sum_{k=1}^{d} f_k (P_k)_{i,j} = \Psi_{i,j} = 0$. Since $\{f_k\}_{k=1}^{d}$ is linearly independent, it follows that $(P_k)_{i,j} = 0$ for all $k \in [d]$ and so $(P - \theta P_k)_{i,j} = 0$. Now suppose that $(P)_{i,j} > 0$. By the definition of θ, it follows that $(\theta P_k)_{i,j} \leq (P)_{i,j}$. Thus, F is a well defined HMM.

Observe that span $\{P - \theta P_k \mid k \in [d]\} \subseteq$ span $\{P_k \mid k \in [d]\}$. The opposite inclusion follows from the fact that, by Equation (1), we have $P \in$ span $\{P - \theta P_k \mid k = 1, \ldots, d\}$. Thus, by Lemma 18,

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

Hence, the proposition follows from Lemma 10. ◀

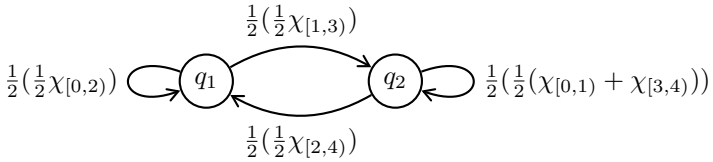
Now we can prove Theorem 12:

Proof of Theorem 12. Suppose the HMM $C = (Q, \Sigma, \Psi)$ is over the linearly decomposable profile language Γ. Let $\Gamma_0 = \{\gamma_1, \ldots, \gamma_n\}$ be the set of profiles appearing in the description of Ψ. From the description of Ψ as a matrix of coefficient-profile pairs, we can easily compute matrices $P'_1, \ldots, P'_n \in \mathbb{Q}^{|Q|\times|Q|}$ such that $\Psi = \sum_{i=1}^{n} [\![\gamma_i]\!] P'_i$. Since Γ is linearly decomposable, one can compute in polynomial time a subset $\{\beta_1, \ldots, \beta_d\} \subseteq \Gamma_0$ such that $[\![\{\beta_1, \ldots, \beta_d\}]\!]$ is linearly independent and also a set of coefficients $b_{i,k}$ such that $[\![\gamma_i]\!] = \sum_{k=1}^{d} b_{i,k} [\![\beta_k]\!]$ for all $i \in [n]$. Hence:

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

Setting $P_k = \sum_{i=1}^{n} b_{i,k} P'_i$ for all $k \in [d]$, we thus obtain the independent functional decomposition $\Psi = \sum_{k=1}^{d} [\![\beta_k]\!] P_k$. Now it is straightforward to compute the finite-observation HMM F from Proposition 21 in polynomial time, thus reducing the equivalence problem in C to the equivalence problem in the finite-observation HMM F. By Proposition 8 the theorem follows. ◀

▶ **Example 22.** We illustrate aspects of the proof of Theorem 12 using the HMM:



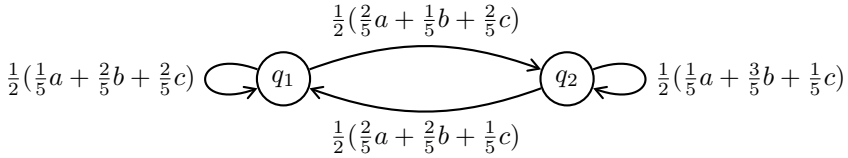
Noting that $\frac{1}{2}(\chi_{[0,1)}+\chi_{[3,4)}) = \frac{1}{2}\chi_{[0,2)} - \frac{1}{2}\chi_{[1,3)} + \frac{1}{2}\chi_{[2,4)}$ and the set $\{\frac{1}{2}\chi_{[0,2)}, \frac{1}{2}\chi_{[1,3)}, \frac{1}{2}\chi_{[2,4)}\}$ is linearly independent we obtain the independent functional decomposition

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

According to Proposition 21, $P = \left(\begin{smallmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\end{smallmatrix}\right)$. Further we compute $\theta = \frac{1}{2}$ and $d - \theta = \frac{5}{2}$ and

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

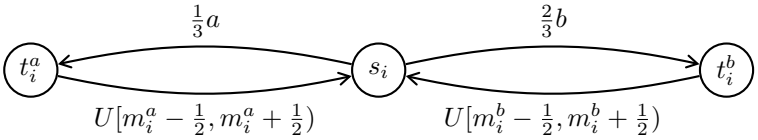
It follows that any initial distributions π_1 and π_2 are equivalent in (Q, Σ, Ψ) if and only if they are equivalent in the following HMM:



For any initial distributions $\pi_1, \pi_2 \in \mathbb{Q}^2$ this can be checked with Proposition 8. (In this example $\pi_1 \equiv \pi_2$ holds only if $\pi_1 = \pi_2$.) ◀

▶ **Example 23.** We also discuss an example, inspired from [6], where HMM non-equivalence means susceptibility to timing attacks, and HMM equivalence means immunity to such attacks. Consider a system that emits two kinds of observations, both visible to an attacker: a function to be executed (we arbitrarily assume a choice between two functions a and b, and impute a probability distribution between them) and the time it takes to execute that function. An attacker therefore sees a sequence $\ell_1 t_1 \ell_2 t_2 \ldots$, where $\ell_i \in \{a, b\}$ and $t_i \in [0, \infty)$. In [6] the times $t_1, t_2, \ldots$ are all identical and depend only on the secret key held by the system, but we assume in the following that the t_i are drawn from a probability distribution that depends on the function (a or b) and the key. We assume that with key i the execution times have uniform distributions $U[m_i^a - \frac{1}{2}, m_i^a + \frac{1}{2})$ and $U[m_i^b - \frac{1}{2}, m_i^b + \frac{1}{2})$. The situation can then be modelled with the HMM below.[2]

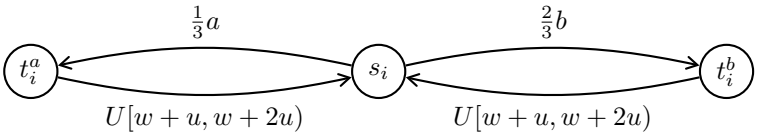
[2] In this case the observation set $\Sigma = [0, \infty) \cup \{a, b\}$ is a disjoint union of topological spaces and there is a natural measure space induced from the Lebesgue measure space on $[0, \infty)$ and a discrete measure on $\{a, b\}$.



A *timing leak* occurs if the attacker can glean the key from the execution times. For example, the attacker can distinguish between keys k_1 and k_2 if and only if states s_1 and s_2 are not equivalent. One can check, using the algorithm we have developed in this section, that s_1 and s_2 are equivalent if and only if $m_1^a = m_2^a$ and $m_1^b = m_2^b$. Moreover, it follows from Section 5 that if instead of $U[m_1^a - \frac{1}{2}, m_1^a + \frac{1}{2})$ and $U[m_2^a - \frac{1}{2}, m_2^a + \frac{1}{2})$ we had two distributions with density functions from $[\![\Gamma_{GEM}]\!]$ with the same mean and the same variance, states s_1, s_2 would still be non-equivalent whenever the two distributions are not identical.

One may try to guard against this timing leak by "*padding*" the execution time, so that the sum of the execution time and an added time is constant (and independent of the key). After the execution of the function, an idling loop would be executed until the worst-case (among all keys) execution time of the functions has been reached or exceeded. Let us call this worst-case execution time $w \in (0, \infty)$. This idling loop would take time $u > 0$ in each iteration, so the total idling time is always an integer multiple of u. It is argued in [6] that this guard is in general ineffective in that the attacker can still glean the execution time modulo u. Therefore, it is suggested in [6] to add, in addition, a time that is uniformly distributed on $[0, u)$.

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



All states s_i are now equivalent, so the key does not leak. ◀

7 Conclusions

We have shown that equivalence of continuous-observation HMMs is decidable in polynomial time, by reduction to the finite-observation case. The crucial insight is that, rather than integrating the density functions, one needs to consider them as elements of a vector space and computationally establish linear (in)dependence of functions. Therefore, our polynomial-time reduction performs symbolic computations on continuous density functions. As a suitable framework for these computations we have introduced the notion of linearly decomposable profile languages, and we have established Γ_{GEM} as such a profile language.

In future work, it would be desirable to extend Γ_{GEM} and/or develop other linear decomposable profile languages, including over sets Σ of observations that are not real numbers. The authors believe that the developed computational framework may be the foundation for further algorithms on continuous-observation HMMs. For example, one may want to compute the total-variation distance of two continuous-observation HMMs. Can Markov chains with continuous emissions be model-checked efficiently?

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. M.S. Alvim, M.E. Andrés, C. Palamidessi, and P. van Rossum. Safe equivalences for security properties. In *Theoretical Computer Science*, pages 55–70. Springer, 2010.
4. C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen. Model-checking algorithms for continuous-time Markov chains. *IEEE Transactions on Software Engineering*, 29(6):524–541, 2003.
5. M.S. Bauer, R. Chadha, and M. Viswanathan. Modular verification of protocol equivalence in the presence of randomness. In *Computer Security – ESORICS 2017*, pages 187–205. Springer, 2017.
6. B.A. Braun, S. Jana, and D. Boneh. Robust and efficient elimination of cache and timing side channels, 2015. `arXiv:1506.00189`.
7. 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.
8. T. Chen, M. Diciolla, M.Z. Kwiatkowska, and A. Mereacre. Time-bounded verification of CTMCs against real-time specifications. In *Proceedings of Formal Modeling and Analysis of Timed Systems (FORMATS)*, volume 6919 of *LNCS*, pages 26–42. Springer, 2011.
9. G.A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1):79–94, 1989.
10. 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.
11. 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.
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. W.K. Grassmann. Finding transient solutions in Markovian event systems through randomization. In *Numerical solution of Markov chains*, pages 357–371, 1991.
16. H. Hermanns, J. Krčál, and J. Křetínský. Probabilistic bisimulation: Naturally on distributions. In *CONCUR 2014 – Concurrency Theory*, pages 249–265. Springer, 2014.
17. S. Kiefer, A.S. Murawski, J. Ouaknine, B. Wachter, and J. Worrell. Language equivalence for probabilistic automata. In *Proceedings of the 23rd International Conference on Computer Aided Verification (CAV)*, volume 6806 of *LNCS*, pages 526–540. Springer, 2011.
18. 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.
19. 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.
20. L.M. Milne-Thomson. *The Calculus of Finite Differences*. Macmillan and Company, 1933.
21. A. Paz. *Introduction to Probabilistic Automata (Computer Science and Applied Mathematics)*. Academic Press, Inc., Orlando, FL, USA, 1971.

22 L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
23 M.P. Schützenberger. On the definition of a family of automata. *Information and Control*, 4(2):245–270, 1961.
24 W.-G. Tzeng. A polynomial-time algorithm for the equivalence of probabilistic automata. *SIAM J. Comput.*, 21(2):216–227, April 1992.

A Missing Proofs

A.1 Proof of Proposition 7 and Lemma 10

▶ **Proposition 7.** *Consider a HMM* (Q, Σ, Ψ). *For any* $\pi_1, \pi_2 \in \mathbb{R}^{|Q|}$ *we have*

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

Unlike in finite probability spaces this fact requires additional assumptions about the space. Therefore we first prove a Lemma that encapsulates these assumptions.

▶ **Lemma 24.** *Let* (Q, Σ, Ψ) *be a HMM and let* π_1, π_2 *be initial distributions. As discussed in the preliminaries,* $(\Sigma, \mathcal{G}, \lambda)$ *is a measure space such that any open set* $E \in \mathcal{G}$ *has non-null measure. Fix* $n \in \mathbb{N}$. *Then,* $(\pi_1 - \pi_2)\Psi(w)\mathbb{1}^T = 0$ *for all* $w \in \Sigma^n$ *if and only if* $(\pi_1 - \pi_2)\int_E \Psi \, d\lambda^n \mathbb{1}^T = 0$ *for all* $E \in \mathcal{G}^n$.

Proof. The forward implication is clear. For the converse, suppose $v \in \Sigma^n$ is such that $(\pi_1 - \pi_2)\Psi(v)\mathbb{1}^T > 0$. Since Ψ is piecewise continuous when restricted to Σ^n, it has a set of continuity C^n as described in the preliminaries and there is a sequence of words $v_k \in C^n$ such that $\lim_{k\to\infty} \Psi(v_k) = \Psi(v)$. As C^n is an open set, there is a sequence of open balls $B(v_k, \epsilon_k) \subseteq C^n$ with $\lim_{k\to\infty} \epsilon_k = 0$. Hence there is $k \in \mathbb{N}$ such that $(\pi_1 - \pi_2)\Psi(w)\mathbb{1}^T > 0$ for all $w \in B(v_k, \epsilon_k)$. By the property of $(\Sigma, \mathcal{G}, \lambda)$ stated in the proposition, we have $\lambda(B(v_k, \epsilon_k)) > 0$ and therefore $(\pi_1 - \pi_2)\int_{B(v_k,\epsilon_k)} \Psi \, d\lambda^n \mathbb{1}^T > 0$. A symmetrical argument can be applied in the case $(\pi_1 - \pi_2)\Psi(v)\mathbb{1}^T < 0$. ◀

We may now prove Proposition 7.

Proof of Proposition 7. We have:

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

where the third equivalence follows from Lemma 24 and is a result of Ψ being piecewise continuous. ◀

Now we prove Lemma 10.

▶ **Lemma 10.** *Let* $C_1 = (Q, \Sigma_1, \Psi_1)$ *and* $C_2 = (Q, \Sigma_2, \Psi_2)$ *be two HMMs with the same state space* Q. *Suppose that* $\operatorname{span}\{\Psi_1(x) \mid x \in \Sigma_1\} \subseteq \operatorname{span}\{\Psi_2(x) \mid x \in \Sigma_2\}$. *Then, for any two initial distributions* π_1 *and* π_2,

$$\pi_1 \equiv_{C_2} \pi_2 \implies \pi_1 \equiv_{C_1} \pi_2.$$

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

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

Thus, $\operatorname{span}\{\Psi_1(w) \mid w \in \Sigma_1^*\} \subseteq \operatorname{span}\{\Psi_2(w) \mid w \in \Sigma_2^*\}$. Therefore, by Proposition 7,

$$\begin{aligned}\pi_1 \equiv_{C_2} \pi_2 &\iff \pi_1 - \pi_2 \perp \operatorname{span}\{\Psi_2(w)\mathbb{1}^T \mid w \in \Sigma_2^*\} \\ &\implies \pi_1 - \pi_2 \perp \operatorname{span}\{\Psi_1(w)\mathbb{1}^T \mid w \in \Sigma_1^*\} \\ &\iff \pi_1 \equiv_{C_1} \pi_2.\end{aligned}$$

◀

A.2 Proof of Proposition 16 and an illustrating example

The main argument for Proposition 16 comes from two Lemmas which we state and prove first.

▶ **Lemma 15.** *Suppose $f_1, \dots, f_n \in \mathcal{L}_1(\Sigma, \lambda)$. Then, the f_i are linearly dependent if and only if all alternant matrices for the f_i are singular.*

Proof. Suppose that the $f_1, \dots, f_n$ are linearly dependent. Then, there exist $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ that are not all 0 such that $\sum_{i=1}^n \lambda_i f_i(x) = 0$ for all $x \in \Sigma$. The same dependence holds for the columns of any alternant matrix for the f_i. This proves the forward implication.

Write $M_{f_1,\dots,f_n}(x_1, \dots, x_n)$ for the alternant matrix generated by the functions $f_1, \dots, f_n$ and input points $x_1, \dots, x_n$ and let $G_{f_1,\dots,f_n} : \Sigma^n \to \mathbb{R}$ be given by $G_{f_1,\dots,f_n}(x_1, \dots, x_n) = \det M_{f_1,\dots,f_n}(x_1, \dots, x_n)$.

For the converse implication, it suffices to show that $G_{f_1,\dots,f_n} = 0$ on Σ^n implies that $\{f_1, \dots, f_n\}$ is linearly dependent. We proceed by induction on the number n of functions. Suppose $n = 1$ with single function f. If for all $x \in \Sigma$, $0 = G_f(x) = f(x)$ then clearly $f = 0$ and $\{0\}$ is a linearly dependent set in any vector space.

Now suppose that for $n \geq 1$ and arbitrary functions $g_1, \dots, g_n : \Sigma \to \mathbb{R}$, $G_{g_1,\dots,g_n} = 0$ implies that $g_1, \dots, g_n$ are linearly dependent. Let $f_1, \dots, f_{n+1} : \Sigma \to \mathbb{R}$ and $x_1, \dots, x_{n+1}$. A Laplace expansion of $G_{f_1,\dots,f_{n+1}}(x_1, \dots, x_{n+1})$ along the first row of $M_{f_1,\dots,f_{n+1}}(x_1, \dots, x_{n+1})$ gives

$$\begin{aligned}G_{f_1,\dots,f_{n+1}}(x_1, \dots, x_{n+1}) = f_1(x_1) &G_{f_2,\dots,f_{n+1}}(x_2, \dots, x_{n+1}) \\ &+ \cdots \\ &+ (-1)^n f_{n+1}(x_1) G_{f_1,\dots,f_n}(x_2, \dots, x_{n+1}).\end{aligned}$$

Suppose $G_{f_1,\dots,f_{n+1}}(x_1, \dots, x_{n+1}) = 0$ holds for all $x_1, \dots, x_n$. We distinguish between two cases.

- Either there exist $x_2, \dots, x_{n+1} \in \Sigma$ such that the cofactors

  $$G_{f_2,\dots,f_{n+1}}(x_2, \dots, x_{n+1}), G_{f_1,f_3,\dots,f_{n+1}}(x_2, \dots, x_{n+1}), \dots, G_{f_1,\dots,f_n}(x_2, \dots, x_{n+1})$$

  are not all 0. This establishes a linear dependence in $f_1, \dots, f_{n+1}$.
- Or all cofactors are 0 for all $x_2, \dots, x_{n+1}$. Then, in particular, $G_{f_2,\dots,f_{n+1}}(x_2, \dots, x_{n+1}) = 0$ for all $x_2, \dots, x_{n+1}$. By the induction hypothesis it follows that the functions $f_2, \dots, f_{n+1}$ are linearly dependent. Hence, so are $f_1, \dots, f_{n+1}$.

In either case it follows that $f_1, \ldots, f_{n+1}$ are linearly dependent. ◀

▶ **Lemma 14.** *Let H be a set of disjoint half open intervals. Suppose that $m_1, \ldots, m_I$ are distinct interval-domain monomials such that* $\operatorname{supp} m_i \in H$ *for all $i \in [I]$. In addition, let $g_1, \ldots, g_J$ and $e_1, \ldots, e_K$ be distinct Gaussian and exponential density functions, respectively. Then, the set $\{m_1, \ldots, m_I, g_1, \ldots, g_J, e_1, \ldots, e_K\}$ is linearly independent.*

Proof. Assume that there is a linear dependence

$$\sum_{i=1}^{I} r_i m_i(x) + \sum_{j=1}^{J} s_j g_j(x) + \sum_{k=1}^{K} t_k e_k(x) = 0 \quad \forall x \in \mathbb{R}.$$

By reordering if necessary, we may assume that the exponential functions $e_1, \ldots, e_K$ have strictly decreasing rates $\lambda_1 > \cdots > \lambda_K$. The function e_K tends to 0 at the slowest rate out of all other functions in the linear dependence and so

$$\lim_{x\to\infty} \frac{1}{e_K(x)} \Big[\sum_{i=1}^{I} r_i m_i(x) + \sum_{j=1}^{J} s_j g_j(x) + \sum_{k=1}^{K} t_k e_k(x) \Big] = t_K \,,$$

which implies that $t_K = 0$. Repeating this argument for decreasing $k \in [K]$ it follows that $t_1 = \cdots = t_K = 0$ and therefore

$$\sum_{i=1}^{I} r_i m_i(x) + \sum_{j=1}^{J} s_j g_j(x) = 0 \quad \forall x \in \mathbb{R}.$$

Suppose the Gaussian functions $g_1, \ldots, g_J$ have mean and standard deviation $\mu_1, \ldots, \mu_J$ and $\sigma_1, \ldots, \sigma_J$, respectively. By defining the ordering $g_i <_{\text{lex}} g_j$ if and only if $\sigma_i < \sigma_j \vee (\sigma_i = \sigma_j \wedge \mu_i < \mu_j)$ we may assume without loss of generality that $g_1 <_{\text{lex}} \cdots <_{\text{lex}} g_J$. It follows that for $1 \le j < J$ the ratio

$$\begin{aligned}
\frac{g_j(x)}{g_J(x)} &= \frac{1}{\sigma_j\sqrt{2\pi}} \exp\Big[-\frac{(x-\mu_j)^2}{2\sigma_j^2} \Big] \Big/ \frac{1}{\sigma_J\sqrt{2\pi}} \exp\Big[-\frac{(x-\mu_J)^2}{2\sigma_J^2} \Big] \\
&= \frac{\sigma_J}{\sigma_j} \exp\Big[\frac{1}{2} \Big(\Big(\frac{1}{\sigma_J^2} - \frac{1}{\sigma_j^2} \Big) x^2 - 2 \Big(\frac{\mu_J}{\sigma_J^2} - \frac{\mu_j}{\sigma_j^2} \Big) x + \Big(\frac{\mu_J^2}{\sigma_J^2} - \frac{\mu_j^2}{\sigma_j^2} \Big) \Big) \Big] \\
&\to 0 \text{ as } x \to \infty \,,
\end{aligned}$$

as $g_j <_{\text{lex}} g_J$ implies that the dominant polynomial coefficient in the exponent is always negative. Any Gaussian density function tends to 0 slower than any interval-domain monomial at $+\infty$, so similarly to the exponential densities,

$$0 = \lim_{x\to\infty} \frac{1}{g_J(x)} \Big[\sum_{i=1}^{I} r_i m_i(x) + \sum_{j=1}^{J} s_j g_j(x) \Big] = s_J \,.$$

By repeating this argument for decreasing $j \in [J]$, we obtain $s_1 = \cdots = s_J = 0$. It remains to show the remaining interval-domain monomials are linearly independent. Since H is finite, all interval-domain monomials on $[a, b)$ have a maximum exponent R. The intervals are disjoint so it suffices to consider a single interval $[a, b)$ and show that the set of monomials $\{x^k \chi_{[a,b)} \mid k \in \{0, \ldots, R\}\}$ is linearly independent. Consider the alternant matrix for $1\chi_{[a,b)}, x\chi_{[a,b)}, \ldots, x^R\chi_{[a,b)}$ and distinct input points $x_1, \ldots, x_{R+1} \in [a, b)$. This matrix is a Vandermonde matrix and by [20, p.9] has full rank. Therefore, by Lemma 15 the set $\{1\chi_{[a,b)}, x\chi_{[a,b)}, \ldots, x^R\chi_{[a,b)}\}$ is linearly independent. ◀

▶ **Proposition 16.** *The profile language Γ_{GEM} is linearly decomposable.*

Proof. Let $\Gamma_0 \subseteq \Gamma_{GEM}$ be a finite set of profiles. Any profile in Γ_0 encodes a linear combination of Gaussians, exponentials and interval-domain monomials. Collect in G_0 and E_0 the profiles of Gaussians and exponentials, respectively, that appear in the description of at least one profile in Γ_0. By sorting the start and end points of the intervals (that appear in the interval-domain monomials) in Γ_0, we compute a finite set H of disjoint intervals such that every interval appearing in Γ_0 is a union of intervals in H. Further, collect in N the set of degrees of monomials in Γ_0. Then we compute a set of profiles M_0 such that $[\![M_0]\!] = \{x^n \chi_{[a,b)} \mid n \in N,\ [a,b) \in H\}$. By Lemma 14, the set $[\![G_0 \cup E_0 \cup M_0]\!]$ is linearly independent. We compute the (unique) coordinates of all functions in $[\![\Gamma_0]\!]$ in terms of that basis.

With these coordinates at hand, we compute a subset $\mathcal{B} \subseteq \Gamma_0$ such that $[\![\mathcal{B}]\!]$ is a basis of span $[\![\Gamma_0]\!]$ as follows. Starting with the $\mathcal{B} = \emptyset$, go through Γ_0 one by one; whenever a profile $\gamma \in \Gamma_0$ is such that $[\![\mathcal{B} \cup \{\gamma\}]\!]$ is linearly independent then add γ to $\mathcal{B}$. The check for linear independence can be performed in terms of the computed coordinates of $[\![\Gamma_0]\!]$ in the basis $[\![G_0 \cup E_0 \cup M_0]\!]$. For the final set $\mathcal{B}$ we have that $[\![\mathcal{B}]\!]$ is a basis of span $[\![\Gamma_0]\!]$. The coefficients that express $[\![\Gamma_0]\!]$ as a linear combination of $[\![\mathcal{B}]\!]$ can be computed similarly. All computations referred to in this proof are polynomial-time. ◀

▶ **Example 25.** We illustrate the proof of Proposition 16 using the HMM discussed in Examples 11, 13 and 20. Recall the encoding of Ψ is given as the matrix

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

with $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \Gamma_{GEM}$ and $[\![\gamma_1]\!] = 2x\chi_{[0,1)}$ and $[\![\gamma_2]\!] = 2(1-x)\chi_{[0,1)}$ and $[\![\gamma_3]\!] = \frac{1}{2}\chi_{[0,2)}$ and $[\![\gamma_4]\!] = \chi_{[0,1)}$. Clearly $\Gamma_0 = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\} \subset \Gamma_{GEM}$. By ordering the start and end points in $[0,2), [1,2)$ we compute $H = \{[0,1), [1,2)\}$. The set of degrees is $N = \{0,1\}$. We then compute the set M_0 of profiles such that $[\![M_0]\!] = \{\chi_{[0,1)}, x\chi_{[0,1)}, \chi_{[1,2)}, x\chi_{[1,2)}\}$ and express $[\![\gamma_1]\!], \ldots, [\![\gamma_4]\!]$ as vectors of coordinates with respect to the basis $[\![M_0]\!]$:

$$\begin{aligned} [\![\gamma_1]\!] &= 2x\chi_{[0,1)} &&= (0,2,0,0) \\ [\![\gamma_2]\!] &= 2(1-x)\chi_{[0,1)} &&= (2,-2,0,0) \\ [\![\gamma_3]\!] &= \frac{1}{2}\chi_{[0,2)} &&= (\frac{1}{2}, 0, \frac{1}{2}, 0) \\ [\![\gamma_4]\!] &= \chi_{[0,1)} &&= (1,0,0,0) \end{aligned}$$

We then compute a basis for this set of vectors: $\{(0,2,0,0), (2,-2,0,0), (\frac{1}{2}, 0, \frac{1}{2}, 0)\}$. This implies that with $\mathcal{B} = \{\gamma_1, \gamma_2, \gamma_3\}$, the set $[\![\mathcal{B}]\!]$ is a basis for span $[\![\Gamma_0]\!]$. Since $(1,0,0,0) = \frac{1}{2}(0,2,0,0) + \frac{1}{2}(2,-2,0,0)$, we express $[\![\gamma_4]\!]$ in terms of $[\![\mathcal{B}]\!]$ by $[\![\gamma_4]\!] = \frac{1}{2}[\![\gamma_1]\!] + \frac{1}{2}[\![\gamma_2]\!]$. ◀