

On the Total Variation Distance of Labelled Markov Chains

Taolue Chen

Middlesex University London

t.l.chen@mdx.ac.uk

Stefan Kiefer

University of Oxford

stekie@cs.ox.ac.uk

Abstract

Labelled Markov chains (LMCs) are widely used in probabilistic verification, speech recognition, computational biology, and many other fields. Checking two LMCs for equivalence is a classical problem subject to extensive studies, while the total variation distance provides a natural measure for the “inequivalence” of two LMCs: it is the maximum difference between probabilities that the LMCs assign to the same event.

In this paper we develop a theory of the total variation distance between two LMCs, with emphasis on the algorithmic aspects: (1) we provide a polynomial-time algorithm for determining whether two LMCs have distance 1, i.e., whether they can almost always be distinguished; (2) we provide an algorithm for approximating the distance with arbitrary precision; and (3) we show that the threshold problem, i.e., whether the distance exceeds a given threshold, is NP-hard and hard for the square-root-sum problem. We also make a connection between the total variation distance and *Bernoulli convolutions*.

Categories and Subject Descriptors G.3 [Probability and Statistics]; D.2.4 [Software/Program Verification]

General Terms Theory

Keywords Labelled Markov Chains, Total Variation Distance

1. Introduction

A (discrete-time, finite-state) *labelled Markov chain (LMC)* has a finite set Q of states and for each state a probability distribution over its outgoing transitions. Each outgoing transition is labelled with a letter from a given finite alphabet Σ , and leads to a target state. Figure 1 depicts two LMCs. The semantics is as follows: The chain starts in a given initial state (or in a random state according to a given initial distribution), picks a random transition according to the state’s distribution over the outgoing transitions, outputs the letter of the transition, moves to the target state, and repeats. In such a way, the chain produces a random infinite sequence of letters, i.e., a random infinite word. We regard this infinite word as “observable” to the environment, whereas the infinite sequence of states remains “internal” to the chain. Formally, an LMC defines a probability space whose samples are infinite words (also called *runs* later) over Σ . In [20], it is classified as a generative model. LMCs

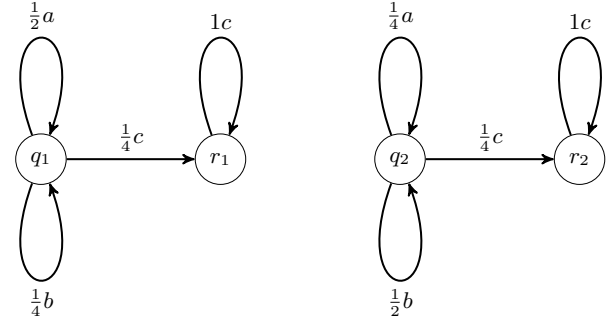


Figure 1. Two LMCs.

appear as *hidden Markov models* in speech recognition and in several areas of computational biology, cf. [15]. LMCs, sometimes in the form of *probabilistic automata* [18], are also fundamental for modelling probabilistic systems.

Checking whether two LMCs (or, similarly, two probabilistic automata) are (*language*) *equivalent* is a classical problem, going back to the seminal work of Schützenberger [19] and Paz [16]. More recently, this problem was revisited, as various verification problems on probabilistic systems can be reduced to it (see, e.g., [13]). As a consequence, efficient polynomial-time algorithms and tools for equivalence checking have been developed [5, 7, 12, 13]. If two systems are found to be *not* equivalent, the question arises on *how different* they are. The *distance* of two LMCs provides a measure for their difference, with the extreme cases being distance 0 for equivalence and distance 1 for (almost-sure) distinguishability. The *total variation distance*, which is a standard distance measure [10] between two probability distributions, yields a natural measure of the distance of two LMCs. Given two probability distributions π_1 and π_2 over the same *countable* set Ω , the total variation distance is defined as

$$d(\pi_1, \pi_2) := \max_{E \subseteq \Omega} |\pi_1(E) - \pi_2(E)|. \quad (1)$$

In words, $d(\pi_1, \pi_2)$ is the largest possible difference between probabilities that π_1 and π_2 assign to the same event. Furthermore, we have $d(\pi_1, \pi_2) = \pi_1(E) - \pi_2(E)$ for

$$E = \{r \in \Omega \mid \pi_1(r) \geq \pi_2(r)\}, \quad (2)$$

so the event E is a maximizer in (1). The total variation distance is—up to a factor of 2—equal to the L_1 -norm of the difference between π_1 and π_2 :

$$2d(\pi_1, \pi_2) = \|\pi_1 - \pi_2\|_1 := \sum_{x \in \Omega} |\pi_1(x) - \pi_2(x)|.$$

When applying the total variance distance to LMCs, it should be emphasized that the sample space $\Omega = \Sigma^\omega$ (i.e., the set of infinite words over Σ) is *uncountable*. Hence the maximum in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CSL-LICS 2014, July 14–18, 2014, Vienna, Austria.
Copyright © 2014 ACM 978-1-4503-2886-9...\$15.00.
<http://dx.doi.org/10.1145/2603088.2603099>

the definition of (total variation) distance needs to be replaced by the supremum. Concretely, assume two LMCs $\mathcal{M}_1, \mathcal{M}_2$ with initial state distributions, the LMCs assign each (measurable) event $E \subseteq \Sigma^\omega$ a probability $\pi_1(E)$ and $\pi_2(E)$, respectively. So the (total variation) distance between $\mathcal{M}_1, \mathcal{M}_2$ is defined as

$$d(\pi_1, \pi_2) := \sup_{E \subseteq \Sigma^\omega} |\pi_1(E) - \pi_2(E)|.$$

It is not clear a priori if a maximizer event exists. We will show later in this paper that it does exist. In particular this means that $d(\pi_1, \pi_2) = 1$ holds if and only if there is an event E with $\pi_1(E) = 1$ and $\pi_2(E) = 0$.

While being an intriguing theoretical question, the study of the distance between LMCs also has practical implications. For instance, in the verification of anonymity properties [12, 13] the following scenario is common: Two users are modelled as LMCs and leave a trace (i.e., emit a run). An evil agent knows the two users, and sees a single trace. The agent wants to find out which of the two users has emitted the trace. Clearly language equivalence (distance 0) of LMCs implies anonymity of the users. If the distance is nonzero, one may ask if the agent can identify the users almost surely. If the distance is 1, the agent succeeds with probability 1, because the agent can define an event E that occurs in the first LMC with probability 1, and in the second one with probability 0; all the agent has to do is to check whether the given run belongs to E . Conversely, if the distance is less than 1, the agent cannot almost-surely distinguish the users. From this point of view, a distance less than 1 is a minimum requirement for some form of user anonymity, which could perhaps be called *deniability*.

Another example is probabilistic model checking where computing the probability of certain events E is of central interest. If the distance between some given LMCs is small (and known or bounded above), computing the probability of E in one of those chains may be enough for obtaining good bounds on the probability of E in the other chains. This may lead to savings in the overall model-checking time.

Main Contributions. In this paper we develop a theory for the total variation distance between two LMCs. We pay special attention to the algorithmic and computational aspects of the problem. We make the following contributions:

- (1) We demonstrate some basic properties of the total variation distance between two LMCs: (a) the supremum in the definition can be “achieved”, and we exhibit a maximizing event, although we show that the maximizing event is *not* ω -regular in general; (b) the distance of two LMCs can be irrational even if all probabilities appearing in their description are rational.
- (2) We study the *qualitative* variant of the distance problem, i.e., to decide whether two LMCs have distance 1 or 0. The distance-0 problem amounts to the language equivalence problem for probabilistic automata, for which a polynomial-time algorithm exists. We provide a polynomial-time algorithm for the distance-1 problem.
- (3) We study the *quantitative* variant of the distance problem. In light of (1), at best one can hope to *approximate* the distance rather than to really *compute* it (at least in the classical complexity theory framework). To this end, we provide an algorithm for approximating the distance with arbitrary precision. We also link the problem to *Bernoulli convolutions* by providing an LMC where the distance of two states of this LMC is related to *Bernoulli convolutions*, thus indicating the intricacy of the distance.
- (4) We study the threshold problem, i.e., to decide whether the distance exceeds a given threshold. While leaving decidability

of the problem open, we show that the problem is both NP-hard and hard for the square-root-sum problem.

Structure of the Paper. In Section 2 we provide technical preliminaries. In Section 3 we give two examples for LMCs and their distances. In Section 4 we discuss two sequences that converge to the distance from below and from above, yielding an approximation algorithm. In Section 5 we show that an event with maximum difference in probabilities always exists, and we exhibit such a “witness” event. In Section 6 we show that the distance can be irrational, and we give lower complexity bounds for the threshold problem. In particular, in Section 6.1 we exhibit an LMC where the distance depends on the probabilities in the LMC in intricate ways, as witnessed by a connection to *Bernoulli convolutions*. In Section 7 we develop a polynomial-time algorithm for deciding whether two LMCs have distance 1. In Section 8 we discuss related work. Finally, in Section 9 we offer some conclusions and highlight open problems. Missing proofs can be found in the full version of the paper [4].

2. Preliminaries

We write $\mathbb{N}$ for the set of nonnegative integers.

Let Q be a finite set. By default we view *vectors*, i.e., elements of $\mathbb{R}^Q$, as row vectors. For a vector $\mu \in [0, 1]^Q$ we write $|\mu| := \sum_{q \in Q} \mu(q)$ for its L_1 -norm. A vector $\mu \in [0, 1]^Q$ is a *distribution* (resp. *subdistribution*) over Q if $|\mu| = 1$ (resp. $|\mu| \leq 1$). For $q \in Q$ we write δ_q for the (*Dirac*) distribution over Q with $\delta_q(q) = 1$ and $\delta_q(r) = 0$ for $r \in Q \setminus \{q\}$. For a subdistribution μ we write $\text{supp}(\mu) = \{q \in Q \mid \mu(q) > 0\}$ for its support. Given two vectors $\mu_1, \mu_2 \in [0, 1]^Q$ we write $\mu_1 \leq \mu_2$ to say that $\mu_1(q) \leq \mu_2(q)$ holds for all $q \in Q$. We view elements of $\mathbb{R}^{Q \times Q}$ as *matrices*. A matrix $M \in [0, 1]^{Q \times Q}$ is called *stochastic* if each row sums up to one, i.e., for all $q \in Q$ we have $\sum_{r \in Q} M(q, r) = 1$.

Definition 1. A labelled (discrete-time, finite-state) Markov chain (LMC) is a tuple $\mathcal{M} = (Q, \Sigma, M)$ where

- Q is a finite set of states,
- Σ is a finite alphabet of labels, and
- $M : \Sigma \rightarrow [0, 1]^{Q \times Q}$ specifies the transitions, so that $\sum_{a \in \Sigma} M(a)$ is a stochastic matrix.

Intuitively, if the LMC is in state q , then with probability $M(a)(q, q')$ it emits a and moves to state q' . For the complexity results of this paper, we assume that all the numbers in the matrices $M(a)$ for $a \in \Sigma$ are rationals given as fractions of integers represented in binary. We extend M to the mapping $M : \Sigma^* \rightarrow [0, 1]^{Q \times Q}$ with $M(a_1 \cdots a_k) = M(a_1) \cdots M(a_k)$ for $a_1, \dots, a_k \in \Sigma$. Intuitively, if the LMC is in state q then with probability $M(w)(q, q')$ it emits the word w and moves (in $|w|$ steps) to state q' .

Fix an LMC $\mathcal{M} = (Q, \Sigma, M)$ for the rest of this section. A run of $\mathcal{M}$ is an infinite sequence $a_1 a_2 \cdots$ with $a_i \in \Sigma$ for all $i \in \mathbb{N}$. We write Σ^ω for the set of runs. For a run $r = a_1 a_2 \cdots$ and $i \in \mathbb{N}$ we write $r_i := a_1 a_2 \cdots a_i$. For a set $W \subseteq \Sigma^*$ of finite words, we define $W\Sigma^\omega := \{w\omega \mid w \in W, \omega \in \Sigma^\omega\} \subseteq \Sigma^\omega$; i.e., the set of runs that have a prefix in W . For $w \in \Sigma^*$ we define $\text{Run}(w) := \{w\}\Sigma^\omega$; i.e., $\text{Run}(w)$ is the set of runs starting with w .

To an (initial) distribution π over Q we associate the probability space $(\Sigma^\omega, \mathcal{F}, \text{Pr}_\pi)$, where $\mathcal{F}$ is the σ -field generated by all *basic cylinders* $\text{Run}(w)$ with $w \in \Sigma^*$, and $\text{Pr}_\pi : \mathcal{F} \rightarrow [0, 1]$ is the unique probability measure such that $\text{Pr}_\pi(\text{Run}(w)) = |\pi M(w)|$. We generalize the definition of Pr_π to subdistributions π in the obvious way, yielding sub-probability measures. An *event* is a measurable set $E \subseteq \Sigma^\omega$. In this paper we consider only measurable subsets of Σ^ω , and when we write $E \subseteq \Sigma^\omega$, the set E is meant to

be measurable. An event is ω -regular, if it is equal to a language accepted by a nondeterministic Büchi automaton. When confusion is unlikely, we may identify the (sub-)distribution π with the induced (sub-)probability measure $\Pr_\pi$; i.e., for events $E \subseteq \Sigma^\omega$ we may write $\pi(E)$ for $\Pr_\pi(E)$. For a distribution π and a word $w \in \Sigma^*$, we write π^w as a shorthand for $\pi M(w)$; intuitively this is the state subdistribution after emitting w . We have $\Pr_\pi(\text{Run}(w)) = |\pi^w|$.

We reserve π, ρ (and $\pi_1, \pi_2, \dots$) for distributions over Q , often viewing π_1, π_2 as given initial distributions. Similarly, we reserve μ, ν for subdistributions over Q . (But note that π^w for $w \in \Sigma^*$ is a subdistribution in general.)

Given two initial distributions π_1, π_2 , we define the (*total variation*) distance between π_1 and π_2 by

$$d(\pi_1, \pi_2) := \sup_{E \subseteq \Sigma^\omega} |\pi_1(E) - \pi_2(E)|.$$

Recall that $E \subseteq \Sigma^\omega$ implicitly means that E is measurable. As $\pi_1(E) - \pi_2(E) = -(\pi_1(\Sigma^\omega \setminus E) - \pi_2(\Sigma^\omega \setminus E))$, we have in fact $d(\pi_1, \pi_2) = \sup_{E \subseteq \Sigma^\omega} (\pi_1(E) - \pi_2(E))$.

Remark 2. One could analogously define the total variation distance between two LMCs $\mathcal{M}_1 = (Q_1, \Sigma, M_1)$ and $\mathcal{M}_2 = (Q_2, \Sigma, M_2)$ with initial distributions π_1 and π_2 over Q_1 and Q_2 , respectively. Our definition is without loss of generality, as one can take the LMC $\mathcal{M} = (Q, \Sigma, M)$ where Q is the disjoint union of Q_1 and Q_2 , and M is defined using M_1 and M_2 in the straightforward manner.

We write $\mu_1 \equiv \mu_2$ to denote that μ_1 and μ_2 are (language) equivalent, i.e., that $|\mu_1^w| = |\mu_2^w|$ holds for all $w \in \Sigma^*$. The following proposition states in particular that equivalence can be decided in polynomial time, and that equivalence and the distance being zero are equivalent.

Proposition 3.

- (a) We have $\pi_1 \equiv \pi_2$ if and only if $d(\pi_1, \pi_2) = 0$.
- (b) One can compute in polynomial time a set $\mathcal{B} \subseteq \mathbb{Q}^{2|Q|}$ of column vectors, with $|\mathcal{B}| \leq 2|Q|$, such that for all subdistributions μ_1, μ_2 we have $\mu_1 \equiv \mu_2$ if and only if $(\mu_1 \mu_2) \cdot b = 0$ holds for all $b \in \mathcal{B}$. Here, $(\mu_1 \mu_2) \in [0, 1]^{2|Q|}$ is the row vector obtained by gluing μ_1, μ_2 together. (Note that $(\mu_1 \mu_2) \cdot b$ is a scalar.)
- (c) We have $\mu_1 \equiv \mu_2$ if and only if $|\mu_1^w| = |\mu_2^w|$ holds for all $w \in \Sigma^*$ with $|w| = 2|Q|$.
- (d) It is decidable in polynomial time whether $\mu_1 \equiv \mu_2$ holds. Hence it is also decidable in polynomial time whether $d(\pi_1, \pi_2) = 0$ holds.

Proposition 3 (a) is immediate from the definitions. Parts (b)–(d) follow from a linear-algebra argument described, e.g., in [7, 16, 19]. We sketch this argument in [4].

3. Examples

We illustrate some phenomena of the distance by two examples. The main observations are that the distance of two LMCs can be irrational (Example 1), and in general, they must be differentiated by events which are *not* ω -regular (Example 1), even if their distance is 1 (Example 2).

3.1 Example 1

Consider the LMCs from Figure 1 on page 1. As discussed in Remark 2, we can equivalently view them as a single LMC. To illustrate the definitions we study the distance between states q_1 and q_2 , or more precisely, between the Dirac distributions δ_{q_1} and δ_{q_2} . Note that we have $\delta_{r_1} \equiv \delta_{r_2}$, as both r_1 and r_2 keep emitting the letter c . On the other hand we have $\delta_{q_1} \not\equiv \delta_{q_2}$ and so $d(\delta_{q_1}, \delta_{q_2}) > 0$. With probability 1, one of the states r_1, r_2 will

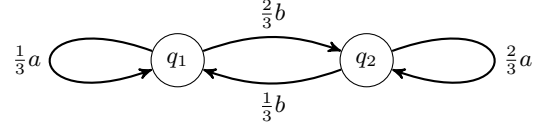


Figure 2. A 2-state LMC. The two states have distance 1.

eventually be reached. So events are characterized by the words over a, b emitted before the infinite c -sequence. More formally, for any event $E \subseteq \Sigma^\omega = \{a, b, c\}^\omega$ one can define $W_E := \{w \in \{a, b\}^* \mid wc^\omega \in E\}$ so that we have

$$\delta_{q_i}(E) = \sum_{w \in W_E} \delta_{q_i}(\{w\}\{c\}^\omega) \quad \text{for } i \in \{1, 2\}.$$

It is easy to see that $\delta_{q_1}(\{a\}\{c\}^\omega) = \frac{1}{8}$ and $\delta_{q_2}(\{a\}\{c\}^\omega) = \frac{1}{16}$. Consider any event E with W_E defined as above. If $a \in W_E$, then $\delta_{q_2}(E) \geq \delta_{q_2}(\{a\}\{c\}^\omega) = \frac{1}{16}$. If $a \notin W_E$, then $\delta_{q_1}(E) \leq 1 - \delta_{q_1}(\{a\}\{c\}^\omega) = \frac{7}{8}$. So for any E we have $\delta_{q_1}(E) - \delta_{q_2}(E) \leq \max\{1 - \frac{1}{16}, \frac{7}{8}\} = \frac{15}{16}$. By symmetry we also have $\delta_{q_2}(E) - \delta_{q_1}(E) \leq \frac{15}{16}$. As E was arbitrary, we have thus shown $d(\delta_{q_1}, \delta_{q_2}) \leq \frac{15}{16} < 1$. We will show in Proposition 12 that we have in fact $d(\delta_{q_1}, \delta_{q_2}) = \sqrt{2}/4$, so distances may be irrational. The proof of Proposition 12 shows that $d(\delta_{q_1}, \delta_{q_2}) = \delta_{q_1}(E) - \delta_{q_2}(E)$ holds for the event

$$E := \{wccc \dots \mid w \in \{a, b\}^*, \#_a(w) \geq \#_b(w)\},$$

where $\#_a(w)$ and $\#_b(w)$ denote the number of occurrences of a and b in w respectively. This may be intuitive as q_1 is more likely to emit a -letters than b -letters, whereas for q_2 it is the opposite. We remark that this event E is not ω -regular, i.e., it cannot be recognized by a Büchi automaton. As a matter of fact, any ω -regular event can only differentiate the two LMCs by a rational number, as the probability of any ω -regular event must be rational.

3.2 Example 2

Consider the LMC in Figure 2. Both states q_1, q_2 can initiate any run $r \in \Sigma^\omega$. Note also that we have $\delta_{q_1}(\{r\}) = \delta_{q_2}(\{r\}) = 0$ for any single run $r \in \Sigma^\omega$. Nevertheless it follows from Theorem 7 that we have $d(\delta_{q_1}, \delta_{q_2}) = 1$. Moreover, Theorem 11 will provide an event E with $\delta_{q_1}(E) = 1$ and $\delta_{q_2}(E) = 0$. Intuitively, such an event could be based on the observation that if q_1 is the initial state, it is more likely after an even number of emitted b -letters to emit another b , whereas if q_2 is the initial state, it is more likely after an even number of emitted b -letters to emit an a -letter. By the law of large numbers, this difference almost surely “shows” in the long run.

In the following we sketch a proof for the fact that no ω -regular event E satisfies both $\delta_{q_1}(E) = 1$ and $\delta_{q_2}(E) = 0$. In fact, we even show that for any ω -regular E with $\delta_{q_1}(E) = 1$ we also have $\delta_{q_2}(E) = 1$. (We omit precise automata-theoretic definitions here, as this argument will play no further role in this paper.) Let E be any ω -regular event. Let R be a deterministic Rabin automaton for E , with initial state r_0 . Let $\mathcal{M}_R$ denote the LMC obtained by taking the cross-product of R and the chain from Figure 2. Let $\delta_{q_1}(E) = 1$. Then all bottom SCCs of $\mathcal{M}_R$ reachable from (r_0, q_1) are accepting. As the qualitative transition structure (i.e., distinguishing only zero and nonzero transition probabilities) is completely symmetric for q_1 and q_2 , it follows that all bottom SCCs of $\mathcal{M}_R$ reachable from (r_0, q_2) are accepting as well. Hence we have $\delta_{q_2}(E) = 1$.

4. An Approximation Algorithm

In this section we define two computable sequences that converge to the distance from below and from above, respectively. This yields an algorithm for approximating the distance with arbitrary precision.

From now on until the end of Section 5 we fix an LMC $\mathcal{M} = (Q, \Sigma, M)$ and (initial) distributions π_1, π_2 . For $w \in \Sigma^*$ we define

$$\begin{aligned} \min(w) &:= \min\{|\pi_1^w|, |\pi_2^w|\} \quad \text{and} \\ \text{con}(w) &:= \max\{|\mu_1| \mid \mu_1 \leq \pi_1^w \wedge \exists \mu_2 \leq \pi_2^w : \mu_1 \equiv \mu_2\}. \end{aligned}$$

For $k \in \mathbb{N}$, we also define $\min(k) := \sum_{w \in \Sigma^k} \min(w)$ and $\text{con}(k) := \sum_{w \in \Sigma^k} \text{con}(w)$. The following proposition lists basic properties of those quantities.

Proposition 4. *Let $w \in \Sigma^*$ and $k \in \mathbb{N}$.*

- (a) *We have $1 \geq \min(w) \geq \text{con}(w) \geq 0$. Hence $\min(k) \geq \text{con}(k)$.*
- (b) *We have $\min(w) \geq \sum_{a \in \Sigma} \min(wa)$ and $\text{con}(w) \leq \sum_{a \in \Sigma} \text{con}(wa)$. Hence we have $\min(k) \geq \min(k+1)$ and $\text{con}(k) \leq \text{con}(k+1)$.*
- (c) *The limits $\min(\infty) := \lim_{i \rightarrow \infty} \min(i)$ and $\text{con}(\infty) := \lim_{i \rightarrow \infty} \text{con}(i)$ exist, and we have $\min(\infty) \geq \text{con}(\infty)$.*

Proof.

- (a) We have $1 \geq |\pi_1^w| \geq \min\{|\pi_1^w|, |\pi_2^w|\} = \min(w)$, hence $1 \geq \min(w)$. Clearly, $\text{con}(w) \geq 0$.
Let μ_1, μ_2 be the subdistributions such that $\text{con}(w) = |\mu_1|$ and $\pi_1^w \geq \mu_1 \equiv \mu_2 \leq \pi_2^w$. Then we have $|\pi_1^w| \geq |\mu_1| = |\mu_2| \leq |\pi_2^w|$, hence $\text{con}(w) = |\mu_1| \leq \min\{|\pi_1^w|, |\pi_2^w|\} = \min(w)$.
- (b) Let $i \in \{1, 2\}$ with $|\pi_i^w| \leq |\pi_{3-i}^w|$. Then we have:

$$\begin{aligned} \min(w) &= \min\{|\pi_1^w|, |\pi_2^w|\} = |\pi_i^w| = |\pi_i M(w)| \\ &= |\pi_i M(w) \sum_{a \in \Sigma} M(a)| \\ &= \sum_{a \in \Sigma} |\pi_i M(wa)| = \sum_{a \in \Sigma} |\pi_i^{wa}| \\ &\geq \sum_{a \in \Sigma} \min\{|\pi_i^{wa}|, |\pi_{3-i}^{wa}|\} = \sum_{a \in \Sigma} \min(wa) \end{aligned}$$

Let μ_1, μ_2 be the subdistributions such that $\text{con}(w) = |\mu_1|$ and $\pi_1^w \geq \mu_1 \equiv \mu_2 \leq \pi_2^w$. It follows that, for all $a \in \Sigma$, we have $\pi_1^{wa} \geq \mu_1 M(a) \equiv \mu_2 M(a) \leq \pi_2^{wa}$, hence $\text{con}(wa) \geq |\mu_1 M(a)|$. So we have:

$$\begin{aligned} \sum_{a \in \Sigma} \text{con}(wa) &\geq \sum_{a \in \Sigma} |\mu_1 M(a)| = |\mu_1| \sum_{a \in \Sigma} M(a) = |\mu_1| \\ &= \text{con}(w). \end{aligned}$$

- (c) Follows from (a) and (b). □

The quantities $\min(k)$ and $\text{con}(k)$ provide lower and upper bounds for the distance:

Proposition 5. *For all $k \in \mathbb{N}$ we have:*

$$1 - \min(k) \leq d(\pi_1, \pi_2) \leq 1 - \text{con}(k).$$

Proof. We show first the lower bound. Let $k \in \mathbb{N}$. Define $W_1 := \{w \in \Sigma^k \mid |\pi_1^w| \geq |\pi_2^w|\}$ and $W_2 := \{w \in \Sigma^k \mid |\pi_1^w| < |\pi_2^w|\}$.

By the definitions we have:

$$\begin{aligned} d(\pi_1, \pi_2) &= \sup_{E \subseteq \Sigma^\omega} (\pi_1(E) - \pi_2(E)) \\ &\geq \pi_1(W_1 \Sigma^\omega) - \pi_2(W_1 \Sigma^\omega) \\ &= 1 - \pi_1(W_2 \Sigma^\omega) - \pi_2(W_1 \Sigma^\omega) \\ &= 1 - \sum_{w \in W_2} |\pi_1^w| - \sum_{w \in W_1} |\pi_2^w| \\ &= 1 - \sum_{w \in \Sigma^k} \min\{|\pi_1^w|, |\pi_2^w|\} \\ &= 1 - \min(k). \end{aligned}$$

Now we show the upper bound. For an event $E \subseteq \Sigma^\omega$ and a word $w \in \Sigma^*$, we denote by $w^{-1}E$ the event $\{u \in \Sigma^\omega \mid wu \in E\}$. For $w \in \Sigma^*$ we write $\mu_1^{(w)}$ and $\mu_2^{(w)}$ to denote subdistributions with $\text{con}(w) = |\mu_1^{(w)}| = |\mu_2^{(w)}|$ and $\pi_1^w \geq \mu_1^{(w)} \equiv \mu_2^{(w)} \leq \pi_2^w$. The following inequalities hold:

$$\begin{aligned} \pi_1^w(w^{-1}E) &= \mu_1^{(w)}(w^{-1}E) + (\pi_1^w - \mu_1^{(w)})(w^{-1}E) \\ &\leq \mu_1^{(w)}(w^{-1}E) + |\pi_1^w| - |\mu_1^{(w)}| \\ \pi_2^w(w^{-1}E) &\geq \mu_2^{(w)}(w^{-1}E) \end{aligned} \quad (3)$$

We have:

$$\begin{aligned} d(\pi_1, \pi_2) &= \sup_{E \subseteq \Sigma^\omega} (\pi_1(E) - \pi_2(E)) \\ &= \sup_{E \subseteq \Sigma^\omega} \sum_{w \in \Sigma^k} \pi_1^w(w^{-1}E) - \pi_2^w(w^{-1}E) \\ &\leq \sup_{E \subseteq \Sigma^\omega} \sum_{w \in \Sigma^k} \mu_1^{(w)}(w^{-1}E) + |\pi_1^w| - |\mu_1^{(w)}| \\ &\quad - \mu_2^{(w)}(w^{-1}E) \quad (\text{by (3)}) \\ &= \sup_{E \subseteq \Sigma^\omega} \sum_{w \in \Sigma^k} |\pi_1^w| - |\mu_1^{(w)}| \quad (\text{as } \mu_1^{(w)} \equiv \mu_2^{(w)}) \\ &= 1 - \sum_{w \in \Sigma^k} |\mu_1^{(w)}| = 1 - \text{con}(k) \end{aligned}$$

□

The lower bound in this proposition follows by considering the event $E_k := W_1 \Sigma^\omega$ (where W_1 is from the proof), which depends only on the length- k prefix of the run. In fact, if we restrict each run to its length- k prefix, we obtain a finite sample space, and the event E_k is the maximizer according to (2) in the introduction. We could define, for each $k \in \mathbb{N}$, a distance $d_k(\pi_1, \pi_2)$ with

$$\begin{aligned} d_k(\pi_1, \pi_2) &= \max_{W \in \Sigma^k} |\pi_1(W \Sigma^\omega) - \pi_2(W \Sigma^\omega)| \\ &= \pi_1(E_k) - \pi_2(E_k) = 1 - \min(k). \end{aligned}$$

Since $d_k(\pi_1, \pi_2) \leq d_{k+1}(\pi_1, \pi_2)$ holds by Proposition 4 (b), there is a limit $\lim_{k \rightarrow \infty} d_k(\pi_1, \pi_2)$, which equals $d(\pi_1, \pi_2)$ (as we will show in Theorem 7). This would offer an alternative but equivalent definition of the distance, which avoids the use of infinite runs by replacing them with increasing prefixes.

By combining Propositions 4 and 5 we obtain

$$1 - \min(\infty) \leq d(\pi_1, \pi_2) \leq 1 - \text{con}(\infty). \quad (4)$$

In the rest of this section we show that those inequalities are in fact equalities.

Recall that for a (random) run $r \in \Sigma^\omega$ we write $r_i \in \Sigma^i$ for the length- i prefix of r . For $i \in \mathbb{N}$, we define the random variable L_i

that assigns to a run $r \in \Sigma^\omega$ the likelihood ratio

$$L_i(r) := |\pi_2^{r_i}| / |\pi_1^{r_i}|.$$

Observe that $L_0(r) = 1$.

Proposition 6. *We have*

$$\begin{aligned} \pi_1 \left(\lim_{i \rightarrow \infty} L_i \text{ exists and is in } [0, \infty) \right) &= 1 \quad \text{and} \\ \pi_2 \left(\lim_{i \rightarrow \infty} 1/L_i \text{ exists and is in } [0, \infty) \right) &= 1. \end{aligned}$$

Proof. We prove only the first equality; the second equality is proved similarly. First we show that the sequence $L_0, L_1, \dots$ is a martingale. Denote by Ex_1 the expectation with respect to π_1 . Let $i \in \mathbb{N}$ and let $w \in \Sigma^i$ with $|\pi_1^w| > 0$. We have:

$$\begin{aligned} \text{Ex}_1(L_{i+1} \mid \text{Run}(w)) &= \sum_{q, q' \in Q} \sum_{a \in \Sigma} \frac{\pi_1^w(q) M(a)(q, q')}{|\pi_1^w|} \cdot \frac{|\pi_2^{wa}|}{|\pi_1^{wa}|} \\ &= \frac{1}{|\pi_1^w|} \sum_{a \in \Sigma} \frac{|\pi_2^{wa}|}{|\pi_1^{wa}|} \underbrace{\sum_{q \in Q} \pi_1^w(q) \sum_{q' \in Q} M(a)(q, q')}_{= |\pi_1^{wa}|} \\ &= \frac{1}{|\pi_1^w|} \sum_{a \in \Sigma} |\pi_2^{wa}| = |\pi_2^w| / |\pi_1^w| = L_i \end{aligned}$$

So $L_0, L_1, \dots$ is a martingale. More precisely, the sequence $L_0, L_1, \dots$ is a nonnegative martingale with $\text{Ex}_1(L_i) = 1$ for all $i \in \mathbb{N}$. So the martingale convergence theorem (more precisely, “Doob’s forward convergence theorem”, see e.g. [21]) applies, and we obtain $\pi_1(\lim_{i \rightarrow \infty} L_i \text{ exists and is finite}) = 1$. $\square$

In the following we may write $\lim_{i \rightarrow \infty} L_i = \infty$ to mean $\lim_{i \rightarrow \infty} 1/L_i = 0$. Define

$$\bar{L} := \lim_{i \rightarrow \infty} L_i \in [0, \infty] \quad (\text{if the limit exists}). \quad (5)$$

The random variable $\bar{L}$ plays a crucial role in the next section and is also used in the proof of the following theorem.

Theorem 7. *We have*

$$1 - \min(\infty) = d(\pi_1, \pi_2) = 1 - \text{con}(\infty).$$

Proof sketch. The proof (see [4]) is somewhat technical and we only give a sketch here. Considering (4) it suffices to show that $\min(\infty) = \text{con}(\infty)$. By Proposition 4 we have $\min(k) \geq \text{con}(k)$ for all k , so loosely speaking we have to show that for “large”¹ k , $\min(k)$ is not much larger than $\text{con}(k)$. We first show that this holds for individual runs started from π_1 ; more precisely, we show for all $\gamma > 0$ that

$$\begin{aligned} \pi_1(\bar{L} > 0) \\ = \pi_1(\bar{L} > 0 \wedge \exists i \in \mathbb{N} : \min(r_i) \leq (1 + \gamma) \text{con}(r_i)). \end{aligned} \quad (6)$$

In words: Conditioned under the event $\{\bar{L} > 0\}$ the probability that eventually $\min(r_i) \leq (1 + \gamma) \text{con}(r_i)$ holds is 1. To show (6) we first show that conditioned under $\{\bar{L} > 0\}$ we have with probability 1 that the distance between the distributions $\pi_1^{r_i}/|\pi_1^{r_i}|$ and $\pi_2^{r_i}/|\pi_2^{r_i}|$ converges to 0. Using the fact that the set of distributions is compact, one can then show (6).

To show that for large k , $\min(k)$ is not much larger than $\text{con}(k)$, we consider a partition $\Sigma^k = W_1 \cup W_2 \cup W_3$. The set W_1 contains the words w with small $|\pi_2^w|/|\pi_1^w|$. So

$\sum_{w \in W_1} \min(w) \leq \sum_{w \in W_1} |\pi_2^w|$ is small. The set W_2 contains the words w with $\min(w) > (1 + \gamma) \text{con}(w)$ and large $|\pi_2^w|/|\pi_1^w|$. Runs with prefixes in W_2 and $\bar{L} = 0$ are unlikely, as $L_k = |\pi_2^w|/|\pi_1^w|$ is large and k is large and the sequence $L_0, L_1, \dots$ converges to $\bar{L}$ by Proposition 6. Runs with prefixes in W_2 and $\bar{L} > 0$ are also unlikely because of (6). So $\sum_{w \in W_2} \min(w) \leq \sum_{w \in W_2} |\pi_1^w|$ is small. Finally, the set W_3 contains the words w with $\min(w) \leq (1 + \gamma) \text{con}(w)$ and large $|\pi_2^w|/|\pi_1^w|$. So $\sum_{w \in W_3} \min(w)$ is (for small γ) not much larger than $\sum_{w \in W_3} \text{con}(w) \leq \text{con}(k)$. By adding the mentioned inequalities we obtain that $\min(k) = \sum_{w \in \Sigma^k} \min(w)$ is not much larger than $\text{con}(k)$. $\square$

Corollary 8. *There is an algorithm that, given $\varepsilon > 0$, computes $a \in \mathbb{Q}$ such that $d(\pi_1, \pi_2) \in [a, a + \varepsilon]$.*

Proof. By Proposition 5 and Theorem 7 the sequences $(1 - \min(k))_{k \in \mathbb{N}}$ and $(1 - \text{con}(k))_{k \in \mathbb{N}}$ converge to $d(\pi_1, \pi_2)$ from below and above, respectively. For each k , the values $\min(k)$ and $\text{con}(k)$ are computable. $\square$

In terms of the complexity of approximating the distance we have the following result:

Proposition 9. *Approximating the distance up to any ε whose size is polynomial in the given LMC is NP-hard with respect to Turing reductions.*

Proof. In [15, Section 6] (see also [5, Theorem 7]), a reduction is given from the *clique decision problem* to show that computing the distance in LMCs is NP-hard. In their reduction the distance is rational and of polynomial size in the input. Using the continued-fraction method (see e.g. Section 2.4 of [8] for an explanation) it follows that a polynomial-time algorithm (if it exists) for approximating the distance can be used to construct a polynomial-time algorithm for computing the distance exactly. Hence the conclusion. $\square$

This NP-hardness result also follows from the proof of [5, Theorem 10].

5. A Maximizing Event

The proof of Theorem 7 does not yield an event E_1 with $\pi_1(E_1) - \pi_2(E_1) = d(\pi_1, \pi_2) \stackrel{\text{def}}{=} \sup_{E \subseteq \Sigma^\omega} |\pi_1(E) - \pi_2(E)|$. In fact, it is not clear a priori whether such an event exists. In this section we exhibit such a “witness” E_1 . It follows that the supremum from the definition of distance is in fact a maximum.

For some intuition recall from (2) in the introduction that in the countable case the event $E_1 = \{r \in \Omega \mid \pi_1(r) \geq \pi_2(r)\}$ is the desired maximizer. In the case of LMCs this does not work, since each individual run may have probability 0 (as, e.g., in Figure 2). However, by rewriting the inequality $\pi_1(r) \geq \pi_2(r)$ as $\pi_2(r)/\pi_1(r) \leq 1$, one is tempted to guess that $\pi_2(r)/\pi_1(r)$ can be replaced by $\bar{L}(r)$ as defined in (5). In the rest of the section we show that this intuition is correct. Define the events

$$E_1 := \{\bar{L} \leq 1\} \quad \text{and} \quad E_2 := \{\bar{L} > 1\}.$$

By Proposition 6 we have

$$\pi_1(E_1) + \pi_1(E_2) = \pi_2(E_1) + \pi_2(E_2) = 1. \quad (7)$$

The following lemma will suffice for showing that E_1 is the desired maximizer.

Lemma 10. *We have $\pi_1(E_2) + \pi_2(E_1) \leq \min(\infty)$.*

¹In the rest of this proof sketch we gloss over the precise meaning of “small”, “not much larger”, etc., and omit the quotation marks.

Proof. Towards a contradiction, suppose that this does not hold. Then there is $k' \in \mathbb{N}$ with $\pi_1(E_2) + \pi_2(E_1) > \min(k')$; hence there is $\gamma > 0$ with

$$\pi_1(E_2) + \pi_2(E_1) > \min(k') + 4\gamma. \quad (8)$$

Choose $\varepsilon \in (0, \gamma]$ small enough so that

$$\pi_1(\bar{L} \in (1, 1 + 2\varepsilon]) \leq \gamma.$$

Using Proposition 6, choose $k \geq k'$ large enough so that we have

$$\begin{aligned} \pi_1(L_k \leq 1 + \varepsilon \wedge \bar{L} > 1 + 2\varepsilon) &\leq \gamma \quad \text{and} \\ \pi_2(L_k > 1 + \varepsilon \wedge \bar{L} \leq 1) &\leq \gamma. \end{aligned} \quad (9)$$

Then we have:

$$\begin{aligned} \pi_1(E_2) &= \pi_1(\bar{L} \in (1, 1 + 2\varepsilon]) + \pi_1(\bar{L} > 1 + 2\varepsilon) \quad (\text{def. of } E_2) \\ &\leq \gamma + \pi_1(\bar{L} > 1 + 2\varepsilon) \quad (\text{choice of } \varepsilon) \\ &= \gamma + \pi_1(L_k > 1 + \varepsilon \wedge \bar{L} > 1 + 2\varepsilon) \\ &\quad + \pi_1(L_k \leq 1 + \varepsilon \wedge \bar{L} > 1 + 2\varepsilon) \\ &\leq \gamma + \pi_1(L_k > 1 + \varepsilon) \\ &\quad + \pi_1(L_k \leq 1 + \varepsilon \wedge \bar{L} > 1 + 2\varepsilon) \\ &\leq 2\gamma + \pi_1(L_k > 1 + \varepsilon) \quad (\text{by (9)}) \end{aligned}$$

Similarly we have:

$$\begin{aligned} \pi_2(E_1) &\leq \pi_2(L_k > 1 + \varepsilon \wedge \bar{L} \leq 1) \\ &\quad + \pi_2(L_k \leq 1 + \varepsilon) \quad (\text{def. of } E_1) \\ &\leq \gamma + \pi_2(L_k \leq 1 + \varepsilon) \quad (\text{by (9)}) \end{aligned}$$

By adding those two inequalities we obtain

$$\begin{aligned} \pi_1(E_2) + \pi_2(E_1) &\leq 3\gamma + \pi_1(L_k > 1 + \varepsilon) + \pi_2(L_k \leq 1 + \varepsilon). \end{aligned} \quad (10)$$

Define the partition of Σ^k in $\Sigma^k = W_1 \cup W_2 \cup W_3$ with

$$\begin{aligned} W_1 &:= \{w \in \Sigma^k \mid |\pi_2^w|/|\pi_1^w| \leq 1\} \\ W_2 &:= \{w \in \Sigma^k \mid 1 < |\pi_2^w|/|\pi_1^w| \leq 1 + \varepsilon\} \\ W_3 &:= \{w \in \Sigma^k \mid 1 + \varepsilon < |\pi_2^w|/|\pi_1^w|\}. \end{aligned}$$

Then we have

$$\begin{aligned} \pi_2(L_k \leq 1) &= \sum_{w \in W_1} |\pi_2^w| \\ \pi_2(1 < L_k \leq 1 + \varepsilon) &= \sum_{w \in W_2} |\pi_2^w| \leq \sum_{w \in W_2} (1 + \varepsilon) |\pi_1^w| \\ \pi_1(L_k > 1 + \varepsilon) &= \sum_{w \in W_3} |\pi_1^w|. \end{aligned}$$

By adding those (in)equalities we obtain

$$\begin{aligned} \pi_1(L_k > 1 + \varepsilon) + \pi_2(L_k \leq 1 + \varepsilon) &\leq (1 + \varepsilon) \min(k) \leq \min(k) + \varepsilon \leq \min(k) + \gamma. \end{aligned}$$

Combining this with (10) yields

$$\pi_1(E_2) + \pi_2(E_1) \leq \min(k) + 4\gamma \leq \min(k') + 4\gamma,$$

thus contradicting (8) as desired. $\square$

Now we can prove that E_1 is the desired maximizing event.

Theorem 11. *We have*

$$d(\pi_1, \pi_2) = \pi_1(E_1) - \pi_2(E_1).$$

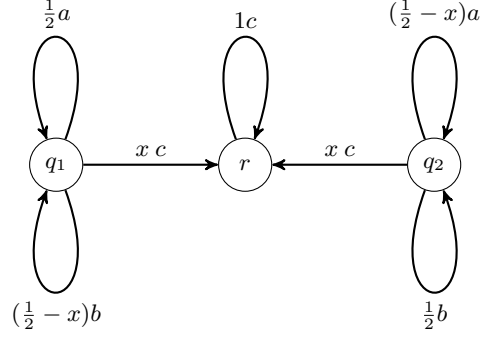


Figure 3. In this LMC, $x \in (0, \frac{1}{2})$ is a parameter. For $x = \frac{1}{4}$ the Dirac distributions δ_{q_1} and δ_{q_2} have distance $\sqrt{2}/4 \notin \mathbb{Q}$.

Proof. We have:

$$\begin{aligned} d(\pi_1, \pi_2) &\geq \pi_1(E_1) - \pi_2(E_1) \quad (\text{definition of distance}) \\ &= 1 - \pi_1(E_2) - \pi_2(E_1) \quad (\text{by (7)}) \\ &\geq 1 - \min(\infty) \quad (\text{Lemma 10}) \\ &= d(\pi_1, \pi_2) \quad (\text{Theorem 7}) \end{aligned}$$

$\square$

6. Irrational Distances and Lower Bounds

The following proposition shows that the distance can be irrational even if all numbers in the description of the LMC are rational.

Proposition 12. *Consider the LMC shown in Figure 3, with parameter $x \in (0, \frac{1}{2})$. We have $d(\delta_{q_1}, \delta_{q_2}) = \frac{1}{2}\sqrt{2x}$.*

We start with a technical lemma.

Lemma 13. *For $y \in [0, \frac{1}{4})$ we have*

$$\sum_{n=0}^{\infty} \binom{2n}{n} y^n = \frac{1}{\sqrt{1-4y}}.$$

Proof. By a binomial series we have:

$$\frac{1}{\sqrt{1-4y}} = (1-4y)^{-1/2} = \sum_{n=0}^{\infty} \binom{-1/2}{n} (-4y)^n$$

By induction on $n \in \mathbb{N}$ one can show that $\binom{-1/2}{n} = \frac{(2n)}{n} \frac{(-1)^n}{4^n}$. The lemma follows. $\square$

Proof of Proposition 12. We write $\pi_1 := \delta_{q_1}$ and $\pi_2 := \delta_{q_2}$. Define $C := \{wccc \dots \mid w \in \{a, b\}^*\} \subseteq \Sigma^\omega$. Clearly we have $\pi_1(C) = \pi_2(C) = 1$. Define

$$\begin{aligned} E_> &:= \{wccc \dots \mid w \in \Sigma^*, \#_a(w) > \#_b(w)\} \quad \text{and} \\ E_< &:= \{wccc \dots \mid w \in \Sigma^*, \#_a(w) < \#_b(w)\}, \end{aligned}$$

where $\#_a(w)$ and $\#_b(w)$ denote the number of occurrences of a resp. b in the word w . The events $E_>, E_<, E_<=$ are defined accordingly.

Recall the event $E_1 = \{\bar{L} \leq 1\} \subseteq \Sigma^\omega$ from Section 5. Using the fact that the LMC in Figure 3 is “deterministic” (i.e., for each $a \in \Sigma$ and $q \in Q$ there is at most one $q' \in Q$ with $M(a)(q, q') > 0$), it is easy to verify that we have $E_1 \cap C = E_>$.

We have:

$$\begin{aligned}
d(\pi_1, \pi_2) &= \pi_1(E_1) - \pi_2(E_1) && \text{(Theorem 11)} \\
&= \pi_1(E_1 \cap C) - \pi_2(E_1 \cap C) && (\text{as } \pi_1(C) = \pi_2(C) = 1) \\
&= \pi_1(E_{\geq}) - \pi_2(E_{\geq}) && (\text{as argued above}) \\
&= \pi_1(E_{\geq}) - \pi_1(E_{\leq}) && (\text{by symmetry of the chain}) \\
&= \pi_1(E_{\geq}) - (1 - \pi_1(E_{>})) && (\text{as } \pi_1(C) = 1) \\
&= 2\pi_1(E_{>}) + \pi_1(E_{=}) - 1 && (\text{by the definitions}). \quad (11)
\end{aligned}$$

The following identity is proved in [11, p.167, (5.20)] and in [14] with a short combinatorial proof:

$$\sum_{m=0}^n \binom{m+n}{m} \left(\frac{1}{2}\right)^m = 2^n \quad \text{for } n \in \mathbb{N}. \quad (12)$$

For $m, n \in \mathbb{N}$ define $E(m, n) := \{wccc \dots \mid w \in \Sigma^*, \#_a(w) = m, \#_b(w) = n\}$. We have:

$$\begin{aligned}
\pi_1(E_{\leq}) &= \sum_{n=0}^{\infty} \sum_{m=0}^n \pi_1(E(m, n)) \\
&= \sum_{n=0}^{\infty} \underbrace{\sum_{m=0}^n \binom{m+n}{m} \left(\frac{1}{2}\right)^m}_{=2^n \text{ by (12)}} \left(\frac{1}{2} - x\right)^n \\
&= \frac{x}{1 - 2\left(\frac{1}{2} - x\right)} = \frac{1}{2}
\end{aligned}$$

So we have $\pi_1(E_{>}) = 1 - \pi_1(E_{\leq}) = \frac{1}{2}$ and hence by (11)

$$d(\pi_1, \pi_2) = \pi_1(E_{=}). \quad (13)$$

We have:

$$\begin{aligned}
\pi_1(E_{=}) &= \sum_{n=0}^{\infty} \pi_1(E(n, n)) \\
&= \sum_{n=0}^{\infty} \binom{2n}{n} \left(\frac{1}{2}\right)^n \left(\frac{1}{2} - x\right)^n x \\
&= \frac{x}{\sqrt{1 - 4\left(\frac{1}{4} - \frac{1}{2}x\right)}} = \frac{1}{2}\sqrt{2x} \quad (\text{by Lemma 13}).
\end{aligned}$$

so the statement follows with (13). $\square$

Note that when $x = \frac{1}{4}$, the LMC shown in Figure 3 is essentially the union of the two LMCs shown in Figure 1. Proposition 12 states that $d(\delta_{q_1}, \delta_{q_2}) = \sqrt{2}/4$, thus substantiating a claim in Section 3.1.

This example suggests that in general it is not obvious what *computing* the distance means, as it may be irrational. Nevertheless it is shown in [15, Section 6] that computing the distance is NP-hard (with respect to Turing reductions). In that reduction the computed LMCs have a rational distance by construction. However, in light of Proposition 12 it may be more natural to study the *threshold-distance* problem defined as follows: Given an LMC, two initial distributions π_1, π_2 , and a threshold $\tau \in [0, 1] \cap \mathbb{Q}$, decide whether $d(\pi_1, \pi_2) \geq \tau$.

By Proposition 9, together with a binary search, the following lower bound follows:

Proposition 14. *The threshold-distance problem is NP-hard with respect to Turing reductions.*

We remark that this can also be done by modifying the reduction from [15], see [4].

In the following we give another lower bound for the threshold-distance problem: the problem is hard for the square-root-sum

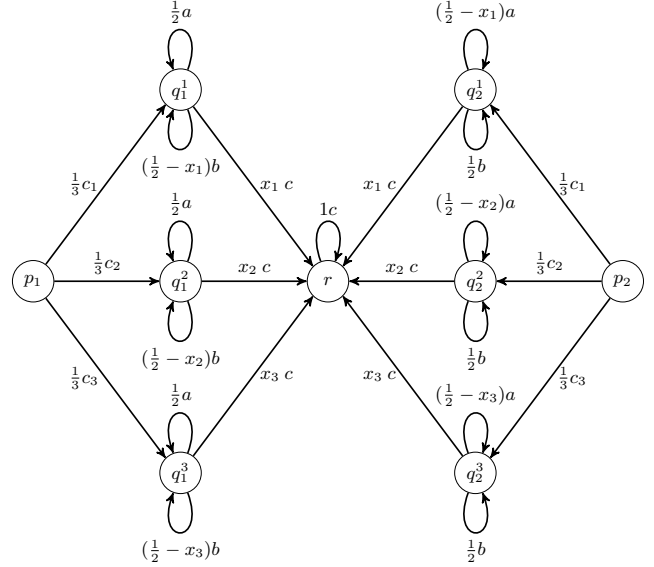


Figure 4. This LMC is obtained by combining the chain from Figure 3 in parallel $n = 3$ times. We have $d(\delta_{p_1}, \delta_{p_2}) = \frac{1}{3} \left(d(\delta_{q_1^1}, \delta_{q_2^1}) + d(\delta_{q_1^2}, \delta_{q_2^2}) + d(\delta_{q_1^3}, \delta_{q_2^3}) \right)$.

problem, as we explain now. Following [1] the *square-root-sum* problem is defined as follows. Given natural numbers $s_1, \dots, s_n \in \mathbb{N}$ and $t \in \mathbb{N}$, decide whether $\sum_{i=1}^n \sqrt{s_i} \geq t$. Membership of square-root-sum in NP has been open since 1976 when Garey, Graham and Johnson [9] showed NP-hardness of the travelling-salesman problem with Euclidean distances, but left membership in NP open. It is known that square-root-sum reduces to PosSLP and hence lies in the 4th level of the counting hierarchy, see [1] and the references therein for more information on square-root-sum, PosSLP, and the counting hierarchy.

We use the LMC from Figure 3 as a “gadget” to prove hardness for the square-root-sum problem:

Theorem 15. *There is a polynomial-time many-one reduction from the square-root-sum problem to the threshold-distance problem.*

Proof sketch. The construction is by taking the LMC from Figure 3 as a gadget, and joining n instances of it in parallel. This is sketched for $n = 3$ in Figure 4. In general we have $\Sigma = \{c_1, \dots, c_n, a, b, c\}$ and $Q = \{p_1, p_2, q_1^1, \dots, q_1^n, q_2^1, \dots, q_2^n, r\}$. Using this construction we have

$$d(\delta_{p_1}, \delta_{p_2}) = \frac{1}{n} \sum_{i=1}^n d(\delta_{q_1^i}, \delta_{q_2^i}). \quad (14)$$

We prove (14) [4]. From the proof of Proposition 12 we know the distances $d(\delta_{q_1^i}, \delta_{q_2^i})$ and the corresponding maximizing events. The proof is completed by suitably choosing the x_i and the threshold τ , see [4]. $\square$

6.1 Bernoulli Convolutions

In this section, we establish another “lower bound” by demonstrating a link to Bernoulli convolutions. Consider the LMC in Figure 5 which has two parameters: $\theta > 1$ and $x \in [-\frac{1}{2}, \frac{1}{2}]$. For each $\theta > 1$, denote by $d_\theta : [-\frac{1}{2}, \frac{1}{2}] \rightarrow [0, 1]$ the function such that $d_\theta(x)$ is the distance between states p_1 and p_2 in the chain with parameters θ and x . Using the Banach fixed-point theorem one can show (see [4]):

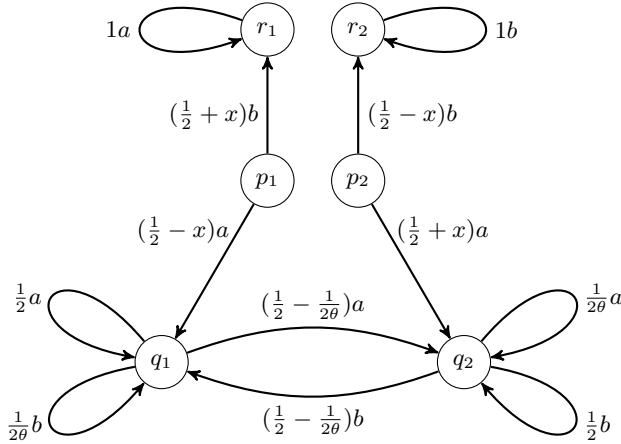


Figure 5. The distance between state p_1, p_2 depends on a Bernoulli-convolution.

Proposition 16. For all $\theta > 1$ we have $d_\theta(x) = \frac{1}{2} + \frac{1}{2}f_\theta(x)$ for the unique function $f_\theta : \mathbb{R} \rightarrow \mathbb{R}$ with

$$f_\theta(x) = \begin{cases} -2x & x \leq -\frac{1}{2} \\ \frac{1}{2\theta}f_\theta(\theta x - (\frac{1}{2}\theta - \frac{1}{2})) & x \in [-\frac{1}{2}, \frac{1}{2}] \\ +\frac{1}{2\theta}f_\theta(\theta x + (\frac{1}{2}\theta - \frac{1}{2})) & x \geq +\frac{1}{2} \end{cases}$$

It follows that the derivative of f_θ must satisfy

$$f'_\theta(x) = \begin{cases} -2 & x \leq -\frac{1}{2} \\ \frac{1}{2}f'_\theta(\theta x - (\frac{1}{2}\theta - \frac{1}{2})) & x \in [-\frac{1}{2}, \frac{1}{2}] \\ +\frac{1}{2}f'_\theta(\theta x + (\frac{1}{2}\theta - \frac{1}{2})) & x \geq +\frac{1}{2} \end{cases} \quad (15)$$

Again, one can use the Banach fixed-point theorem to show that the solution f'_θ is unique.

The functional equation (15) is known from the study of *Bernoulli convolutions*, see [17] for a survey and [2, Chapter 5] for a gentle introduction. In this field the solution of (15) occurs (translated and rescaled) as the cumulative distribution function of the random variable $\sum_{i=0}^{\infty} X_i/\theta^i$, where the X_i are random variables that take on -1 and $+1$ with probability $\frac{1}{2}$ each. Bernoulli convolutions have been studied since the 1930s. It is known that the solutions of (15) are either absolutely continuous or singular on $[-\frac{1}{2}, \frac{1}{2}]$, depending on θ . For $\theta > 2$ they are singular; in fact, for $\theta = 3$ the function is the (ternary) Cantor function. For $\theta = 2$ we have $f'_2(x) = 4x$ for $x \in [-\frac{1}{2}, \frac{1}{2}]$. Erdős showed that if θ is a *Pisot number*², then f'_θ is singular. However, for almost all $\theta \in (1, 2]$ the function f'_θ is absolutely continuous. It is open, e.g., for $\theta = 3/2$ whether f'_θ is absolutely continuous or purely singular.

We conclude from this relation to Bernoulli convolutions that the distance can depend on the probabilities in the LMC in intricate ways.

² A Pisot number is a real algebraic integer greater than 1 such that all its Galois conjugates are less than 1 in absolute value. The smallest Pisot number (≈ 1.3247) is the real root of $x^3 - x - 1$. Another one is the golden ratio $(\sqrt{5} + 1)/2 \approx 1.6180$.

7. The Distance-1 Problem

The *distance-1 problem* asks whether $d(\pi_1, \pi_2) = 1$ holds for a given LMC and two distributions π_1, π_2 . For the rest of the section we fix an LMC $\mathcal{M} = (Q, \Sigma, M)$ and initial distributions π_1, π_2 . Recall from Proposition 3 that $d(\pi_1, \pi_2) = 0$ is equivalent to $\pi_1 \equiv \pi_2$, and that the latter problem, language equivalence, is known to be decidable in polynomial time [13]. In this section we show that the distance-1 problem can also be decided in polynomial time. The algorithm and its correctness argument are much more subtle. The following proposition provides a characterisation of the case $d(\pi_1, \pi_2) < 1$.

Proposition 17. We have $d(\pi_1, \pi_2) < 1$ if and only if there are $w \in \Sigma^*$ and subdistributions μ_1, μ_2 with $\mu_1 \leq \pi_1^w$ and $\mu_2 \leq \pi_2^w$ and $\mu_1 \equiv \mu_2$ and $|\mu_1| = |\mu_2| > 0$.

Note that $\mu_1 \equiv \mu_2$ implies $|\mu_1| = |\mu_2|$. Proposition 17 follows immediately from Theorem 7.

Given π_1, π_2 and a word $w \in \Sigma^*$ one can compute π_1^w and π_2^w in polynomial time. Consider the following condition on w :

$$\exists \mu_1, \mu_2 : \mu_1 \leq \pi_1^w \text{ and } \mu_2 \leq \pi_2^w \text{ and } \mu_1 \equiv \mu_2 \text{ and } |\mu_1| > 0. \quad (16)$$

By Proposition 3 (b), (16) amounts to a feasibility test of a linear program, and hence can be decided in polynomial time. By Proposition 17 we have $d(\pi_1, \pi_2) < 1$ if and only if there is $w \in \Sigma^*$ such that (16) holds.

For notational convenience we write $\text{supp}(w)$ for the pair $(\text{supp}(\pi_1^w), \text{supp}(\pi_2^w))$ in the following. The condition (16) on w is in fact only a condition on $\text{supp}(w)$, as $\mu_1 \equiv \mu_2$ implies $a\mu_1 \equiv a\mu_2$ for all $a \in [0, \infty)$. So (16) can be rephrased as

$$\exists \mu_1, \mu_2 : \text{supp}(\mu_1) \subseteq \text{supp}(\pi_1^w) \text{ and } \text{supp}(\mu_2) \subseteq \text{supp}(\pi_2^w) \text{ and } \mu_1 \equiv \mu_2 \text{ and } |\mu_1| > 0. \quad (17)$$

Moreover, for any two words $w, w' \in \Sigma^*$ with $\text{supp}(w) = \text{supp}(w')$ we have $\text{supp}(wa) = \text{supp}(w'a)$ for all $a \in \Sigma$. This implies

$$\{\text{supp}(w) \mid w \in \Sigma^*\} = \{\text{supp}(w) \mid w \in \Sigma^*, |w| \leq 2^{|Q|}\}.$$

This suggests the following nondeterministic algorithm for checking whether $d(\pi_1, \pi_2) < 1$ holds: compute $\text{supp}(w)$ for a guessed word $w \in \Sigma^*$ with $|w| \leq 2^{|Q|}$ and check (17) for feasibility. Note that w may have exponential length but need not be stored as a whole. This results in a PSPACE algorithm.

In the following, we give a polynomial-time algorithm, which is based on further properties of the distance.

Given subdistributions μ_1, μ_2 with $|\mu_1|, |\mu_2| > 0$ we define the following relation:

$$\mu_1 \sim \mu_2 \iff d\left(\frac{\mu_1}{|\mu_1|}, \frac{\mu_2}{|\mu_2|}\right) < 1$$

Note that $\frac{\mu_1}{|\mu_1|}$ and $\frac{\mu_2}{|\mu_2|}$ are distributions. We have that $\mu_1 \equiv \mu_2$ implies $\mu_1 \sim \mu_2$. The relation $\sim$ is reflexive, symmetric, but in general not transitive. We observe:

Proposition 18. Let $\mu_1 \sim \mu_2$. Let $w \in \Sigma^*$ such that $\text{supp}(\pi_1^w) \supseteq \text{supp}(\mu_1)$ and $\text{supp}(\pi_2^w) \supseteq \text{supp}(\mu_2)$. Then $d(\pi_1, \pi_2) < 1$.

Proof. Since $\mu_1 \sim \mu_2$, we have $d(\rho_1, \rho_2) < 1$ for the distributions $\rho_1 := \mu_1/|\mu_1|$ and $\rho_2 := \mu_2/|\mu_2|$. By Proposition 17 there is a word $v \in \Sigma^*$ and subdistributions ν_1, ν_2 with $|\nu_1| = |\nu_2| > 0$ and $\rho_1^v \geq \nu_1 \equiv \nu_2 \leq \rho_2^v$. Since $\text{supp}(\pi_i^w) \supseteq \text{supp}(\mu_i) = \text{supp}(\rho_i)$ holds for $i \in \{1, 2\}$, we get $\pi_1^{wv} \geq a\nu_1 \equiv a\nu_2 \leq \pi_2^{wv}$ for some small enough $a > 0$. Using Proposition 17 again it follows that $d(\pi_1, \pi_2) < 1$. $\square$

The following proposition states two structural properties of the relation $\sim$ which can be proved using the fact that $d(\pi_1, \pi_2) = 1$ implies that there is a “maximizing” event E with $\pi_1(E) = 1$ and $\pi_2(E) = 0$, see Theorem 11.

Proposition 19. *We have the following.*

- (a) Let $\mu_1 \equiv \mu_2$. Let $\nu_1 \leq \mu_1$ with $|\nu_1| > 0$. Then $\nu_1 \sim \mu_2$.
- (b) Let $\mu_1 \sim \mu_2$. Then there is $q \in \text{supp}(\mu_1)$ with $\delta_q \sim \mu_2$.

Proof.

- (a) Towards a contradiction suppose that $d(\nu_1/|\nu_1|, \mu_2/|\mu_2|) = 1$. Then by Theorem 11 there is an event $E \subseteq \Sigma^\omega$ with $\frac{\nu_1}{|\nu_1|}(E) = 0$ and $\frac{\mu_2}{|\mu_2|}(E) = 1$, i.e., $\nu_1(E) = 0$ and $\mu_2(E) = |\mu_2|$. We have:

$$\begin{aligned}
|\mu_2| &= \mu_2(E) \\
&= \mu_1(E) && (\text{as } \mu_1 \equiv \mu_2) \\
&= (\mu_1 - \nu_1)(E) + \nu_1(E) && (\text{as } \nu_1 \leq \mu_1) \\
&= (\mu_1 - \nu_1)(E) && (\text{as } \nu_1(E) = 0) \\
&\leq |\mu_1 - \nu_1| \\
&= |\mu_1| - |\nu_1| && (\text{as } \nu_1 \leq \mu_1) \\
&< |\mu_1| && (\text{as } |\nu_1| > 0) \\
&= |\mu_2| && (\text{as } \mu_1 \equiv \mu_2),
\end{aligned}$$

which is a contradiction. Hence $\nu_1 \sim \mu_2$.

- (b) Suppose that for all $q \in \text{supp}(\mu_1)$ we have $\delta_q \not\sim \mu_2$, i.e., $d(\delta_q, \mu_2/|\mu_2|) = 1$. By Theorem 11 for all $q \in \text{supp}(\mu_1)$ there is an event $E_q \subseteq \Sigma^\omega$ with $\delta_q(E_q) = 1$ and $\mu_2/|\mu_2|(E_q) = 0$. Consider the event

$$E := \bigcup_{q \in \text{supp}(\mu_1)} E_q.$$

For all $q \in \text{supp}(\mu_1)$ we have $\delta_q(E) \geq \delta_q(E_q) = 1$, so $\delta_q(E) = 1$. Hence,

$$\mu_1(E) = \sum_{q \in \text{supp}(\mu_1)} \mu_1(q) \delta_q(E) = \sum_{q \in \text{supp}(\mu_1)} \mu_1(q) = |\mu_1|.$$

On the other hand, by a union bound, we have

$$\mu_2(E) \leq \sum_{q \in \text{supp}(\mu_1)} \mu_2(E_q) = 0.$$

If $|\mu_2| > 0$, then by the definition of the distance we have

$$d\left(\frac{\mu_1}{|\mu_1|}, \frac{\mu_2}{|\mu_2|}\right) \geq \frac{\mu_1}{|\mu_1|}(E) - \frac{\mu_2}{|\mu_2|}(E) = 1 - 0 = 1,$$

so $\mu_1 \not\sim \mu_2$. If $|\mu_2| = 0$, then by the definition of $\sim$ it also follows $\mu_1 \not\sim \mu_2$. $\square$

For distributions π_1, π_2 we define a set $R^{\pi_1, \pi_2} \subseteq Q \times Q$:

$$R^{\pi_1, \pi_2} := \{(r_1, r_2) \in Q \times Q \mid \exists w \in \Sigma^* : r_1 \in \text{supp}(\pi_1^w) \text{ and } r_2 \in \text{supp}(\pi_2^w)\}$$

This set can be computed in polynomial time:

Lemma 20. *Let π_1, π_2 be distributions. Define a directed graph G as follows. The vertex set is $Q \times Q$. There is an edge from $(q_1, q_2) \in Q \times Q$ to $(r_1, r_2) \in Q \times Q$ if there is a $\alpha \in \Sigma$ with $M(\alpha)(q_1, r_1) > 0$*

Algorithm 1 Polynomial-time algorithm for deciding the distance-1 problem.

procedure distance1

input: LMC $\mathcal{M} = (Q, \Sigma, M)$

initial distributions $\pi_1, \pi_2 \in [0, 1]^Q$

output: $d(\pi_1, \pi_2) = 1$ or $d(\pi_1, \pi_2) < 1$

compute $\mathcal{B} \subseteq \mathbb{Q}^{2|Q|}$ from Proposition 3 (b)

compute R^{π_1, π_2} by graph reachability (Lemma 20)

for $r_1 \in Q$ do

if there exist subdistributions μ_1, μ_2 with

$r_1 \in \text{supp}(\mu_1)$ and $\text{supp}(\mu_2) \subseteq R_{r_1}^{\pi_1, \pi_2}$
and $\forall b \in \mathcal{B} : (\mu_1 \mu_2) \cdot b = 0$

(* this can be decided using linear programming *)

then return “ $d(\pi_1, \pi_2) < 1$ ”

fi

od

return “ $d(\pi_1, \pi_2) = 1$ ”

and $M(a)(q_2, r_2) > 0$. Then we have:

$$\begin{aligned}
R^{\pi_1, \pi_2} \\
&= \{(r_1, r_2) \in Q \times Q \mid \exists q_1 \in \text{supp}(\pi_1) \exists q_2 \in \text{supp}(\pi_2) : \\
&\quad (r_1, r_2) \text{ is reachable from } (q_1, q_2) \text{ in } G\}
\end{aligned}$$

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

The proof of Lemma 20 is straightforward by induction. For $r_1 \in Q$ we define the projection $R_{r_1}^{\pi_1, \pi_2} := \{r_2 \in Q \mid (r_1, r_2) \in R^{\pi_1, \pi_2}\}$. We are ready to show the main theorem of the section:

Theorem 21. *Let π_1, π_2 be distributions. Then $d(\pi_1, \pi_2) < 1$ holds if and only if there are $r_1 \in Q$ and subdistributions μ_1, μ_2 such that*

$$\mu_1 \equiv \mu_2 \text{ and } r_1 \in \text{supp}(\mu_1) \text{ and } \text{supp}(\mu_2) \subseteq R_{r_1}^{\pi_1, \pi_2}.$$

Proof. ($\implies$) Let $d(\pi_1, \pi_2) < 1$. Then by Proposition 17 there are $w \in \Sigma^*$ and subdistributions μ_1, μ_2 with $\mu_1 \leq \pi_1^w$ and $\mu_2 \leq \pi_2^w$ and $\mu_1 \equiv \mu_2$ and $|\mu_1| = |\mu_2| > 0$. By the definition of R^{π_1, π_2} we have $(r_1, r_2) \in R^{\pi_1, \pi_2}$ for all $r_1 \in \text{supp}(\mu_1)$ and all $r_2 \in \text{supp}(\mu_2)$. Choose any $r_1 \in \text{supp}(\mu_1)$. Then $\text{supp}(\mu_2) \subseteq R_{r_1}^{\pi_1, \pi_2}$.

($\impliedby$) For the converse, let $r_1 \in Q$ and μ_1, μ_2 be subdistributions such that

$$\mu_1 \equiv \mu_2 \text{ and } r_1 \in \text{supp}(\mu_1) \text{ and } \text{supp}(\mu_2) \subseteq R_{r_1}^{\pi_1, \pi_2}.$$

By Proposition 19 (a) we have $\delta_{r_1} \sim \mu_2$. Hence, by Proposition 19 (b) there is $r_2 \in \text{supp}(\mu_2)$ with $\delta_{r_1} \sim \delta_{r_2}$. Since $(r_1, r_2) \in R^{\pi_1, \pi_2}$, we have by the definition of R^{π_1, π_2} that there is $w \in \Sigma^*$ with $r_1 \in \text{supp}(\pi_1^w)$ and $r_2 \in \text{supp}(\pi_2^w)$. As $\delta_{r_1} \sim \delta_{r_2}$, Proposition 18 implies $d(\pi_1, \pi_2) < 1$. $\square$

We highlight the algorithmic nature of Theorem 21 in Algorithm 1.

8. Related Work

Two LMCs have distance 0 if and only if they are language equivalent. We have discussed works on language equivalence in the introduction. The papers [15] and [5] are closest to ours. They investigate the L_p -distance between two hidden Markov models [15] and two probabilistic automata [5]. Those models are similar to ours. The main difference is that in their models no letters are emitted once a special *end state* is reached, and the transition structure of the chains guarantees that an end state is eventually reached with probability 1. (Our model is more general, as one can make the

LMC emit an infinite sequence of a special “end letter” once the end state is reached.) So those models induce a probability distribution over Σ^* , which makes the sample space countable. As mentioned in the introduction, the L_1 -distance is then twice the total variation distance, so the hardness results from [5, 15] become available: it is NP-hard to “compute” the L_1 - and L_∞ -distance [15] and the L_p -distance for odd p [5], but recall our discussion after Proposition 12 about irrational distances. We note that the example from Figure 3, showing the existence of irrational distances, can be easily framed in their models. It is also shown in [5] that it is NP-hard to approximate the L_1 -distance within an additive error, and that the L_p -distance can be computed in polynomial time for even p .

The total variation distance in LMCs is considered in [3], where the authors give an upper bound on the total variation distance in terms of the *bisimilarity pseudometric* defined in [6]. Bisimilarity is a “structural” (i.e., based on the emitted letters and the states) notion of equivalence of LMCs, whereas language equivalence is purely “semantical” (based only on the emitted letters). Accordingly, the bisimilarity pseudometric defines a *branching-time* distance while the total variance distance defines a *linear-time* distance. The authors of [3] prove a quantitative analogue of the fact that bisimilarity implies language equivalence: they prove that the bisimilarity pseudometric, which can be computed in polynomial time [3], is an upper bound on the total variation distance which we discuss here.

9. Conclusions and Open Problems

In this paper we have developed a theory of the total variation distance between two LMCs. Two important theoretical results of this paper are summarized as:

- (1) By considering longer and longer prefix words, one can define two sequences $(1 - \min(i))_{i \in \mathbb{N}}$ and $(1 - \text{con}(i))_{i \in \mathbb{N}}$ that converge to the distance from below and above, respectively.
- (2) Using the martingale convergence theorem one can show that there is always a maximizing event, and we have explicitly exhibited one.

These results have algorithmic consequences. Our main algorithmic result is a procedure that decides the distance-1 problem in polynomial time. The result (1) also leads to an algorithm for approximating the distance with arbitrary precision. We have also shown that the distance can be irrational, and we have given lower complexity bounds for the threshold-distance problem: it is NP-hard and hard for the square-root-sum problem.

The complexity and even the decidability of the threshold-distance problem are open problems. A theoretical question is whether the distance is always algebraic. We have established a connection to Bernoulli convolutions, the long history of which may hint at the difficulty of solving the mentioned open problems.

Acknowledgements.

We would like to thank Christoph Haase for pointing us to [11], James Worrell for valuable discussions, and anonymous referees for helpful comments. Stefan Kiefer is supported by a Royal Society University Research Fellowship.

References

- [1] E. Allender, P. Bürgisser, J. Kjeldgaard-Pedersen, and P. B. Miltersen. On the complexity of numerical analysis. *SIAM J. Comput.*, 38(5):1987–2006, 2009.
- [2] D. Bailey, J. Borwein, N. Calkin, R. Girgensohn, D. Luke, and V. Moll. *Experimental Mathematics in Action*. Wellesley, 2007.
- [3] D. Chen, F. van Breugel, and J. Worrell. On the complexity of computing probabilistic bisimilarity. In L. Birkedal, editor, *FoSSaCS*, volume 7213 of *Lecture Notes in Computer Science*, pages 437–451. Springer, 2012. ISBN 978-3-642-28728-2.
- [4] T. Chen and S. Kiefer. On the total variation distance of labelled Markov chains. *CoRR*, abs/1405.2852, 2014. Available at <http://arxiv.org/abs/1405.2852>.
- [5] 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.
- [6] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Metrics for labelled Markov processes. *Theoretical Computer Science*, 318(3):323–354, 2004.
- [7] 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.
- [8] K. Etessami and M. Yannakakis. On the complexity of nash equilibria and other fixed points. *SIAM J. Comput.*, 39(6):2531–2597, 2010.
- [9] M. Garey, R. Graham, and D. Johnson. Some NP-complete geometric problems. In *STOC*, pages 10–22. ACM, 1976.
- [10] A. Gibbs and F. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.
- [11] R. Graham, D. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, second edition, 1989.
- [12] S. Kiefer, A. Murawski, J. Ouaknine, B. Wachter, and J. Worrell. Language equivalence for probabilistic automata. In *CAV*, volume 6806 of *LNCS*, pages 526–540, 2011.
- [13] S. Kiefer, A. Murawski, J. Ouaknine, B. Wachter, and J. Worrell. On the complexity of equivalence and minimisation for Q-weighted automata. *Logical Methods in Computer Science (LMCS)*, 9(1:8):1–22, 2013.
- [14] T. Lengyel. A combinatorial identity and the world series. *SIAM Review*, 35(2):294–297, 1993.
- [15] R. Lyngsø and C. Pedersen. The consensus string problem and the complexity of comparing hidden markov models. *J. Comput. Syst. Sci.*, 65(3):545–569, 2002.
- [16] A. Paz. *Introduction to Probabilistic Automata*. Academic Press, 1971.
- [17] Y. Peres, W. Schlag, and B. Solomyak. Sixty years of Bernoulli convolutions. In *Fractal Geometry and Stochastics II*, volume 46 of *Progress in Probability*, pages 39–65. Birkhäuser Basel, 2000.
- [18] M. O. Rabin. Probabilistic automata. *Information and Control*, 6(3):230–245, 1963.
- [19] M.-P. Schützenberger. On the definition of a family of automata. *Inf. and Control*, 4:245–270, 1961.
- [20] R. van Glabbeek, S. Smolka, and B. Steffen. Reactive, generative and stratified models of probabilistic processes. *Inf. Comput.*, 121(1):59–80, 1995.
- [21] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.