Continuity Properties of Distances for Markov Processes (With Proof Appendix)

Manfred Jaeger[1], Hua Mao[2], Kim G. Larsen[1], and Radu Mardare[1]

[1] Department of Computer Science, Aalborg University, Denmark
{jaeger,kgl,mardare}@cs.aau.dk
[2] College of Computer Science, Sichuan University, PR China
maohuamh@gmail.com

Abstract. In this paper we investigate distance functions on finite state Markov processes that measure the behavioural similarity of non-bisimilar processes. We consider both probabilistic bisimilarity metrics, and trace-based distances derived from standard Lp and Kullback-Leibler distances. Two desirable continuity properties for such distances are identified. We then establish a number of results that show that these two properties are in conflict, and not simultaneously fulfilled by any of our candidate natural distance functions. An impossibility result is derived that explains to some extent the fundamental difficulty we encounter.

1 Introduction

Markov processes are widely used as formal system models in the presence of uncertainty. In the formal analysis of such models, notions of equivalence traditionally play a key role [10]. However, there is an increasing interest in approximate models, such as simplified models obtained by model abstraction, or models that are automatically learned by statistical inference from empirical data [14, 11]. When analysing the relationship between a true model and its approximation, then equivalence clearly is too strong a criterion. Therefore, concepts of approximate equivalence that generalize probabilistic bisimulation equivalence via the introduction of bisimulation distances have received some attention [6, 20, 19, 3, 1].

It turns out however, that some of these distances violate some natural properties one would expect from a distance function that in a meaningful sense measures the quality of approximation. As an example, consider the automaton $\mathcal{M}_\epsilon$ shown in Figure 1 representing a process where a biased ($\epsilon \neq 0$) or unbiased ($\epsilon = 0$) coin is tossed repeatedly. For a small $\epsilon > 0$ the model $\mathcal{M}_\epsilon$ would be considered a good approximation of the model $\mathcal{M}_0$, and if a distance measure d represents quality of approximation, then $d(\mathcal{M}_0, \mathcal{M}_\epsilon)$ should go to zero as $\epsilon \to 0$. This property, which we will formalize as *parameter continuity*, is not satisfied by the original bisimulation distances (though it turns out to be satisfied by the discounted versions of these distances).

Parameter continuity is not the only requirement we have on a distance function. It should also be the case that when one model is a good approximation of another model according to a given distance function, then some upper bounds are implied on the error incurred by using the approximation instead of the real model. We can, thus, formulate two high-level objectives for the design of a distance function:

O1 If $(\mathcal{M}_n)_n$ is a sequence of approximations for a target model $\mathcal{M}$, and for increasing n, $\mathcal{M}_n$ is obtained by applying an increasing amount of resources to obtain a good approximation, then $d(\mathcal{M}, \mathcal{M}_n) \to 0$.

O2 In a particular use scenario for an approximate model $\mathcal{M}'$, an upper distance bound $d(\mathcal{M}, \mathcal{M}') < \delta$ between $\mathcal{M}'$ and the correct model $\mathcal{M}$ should imply an upper bound on the error, or loss, incurred when using $\mathcal{M}'$ instead of $\mathcal{M}$ in the given scenario.

We here have formulated these two objectives in a deliberately vague manner in order to emphasize that they can give rise to a variety of more concrete, formal conditions. One aspect of objective O1 will be captured by the parameter continuity condition illustrated by Figure 1, and formally defined in Section 4.1 below. Parameter continuity matches the informal description of O1 in the sense that if the correct model is $\mathcal{M}_0$, then models obtained by an increasing amount of approximation effort (e.g., learned or constructed from an increasing amount of empirical data) will be of the form $\mathcal{M}_\epsilon$ with decreasing ϵ.

Objective O2 was the main design criterion in the development of the probabilistic bisimulation metrics: a bound on the probabilistic bisimulation distance implies the same bound on the difference in probability for all properties definable in certain formal languages. We follow the same approach, and partly capture the broad objective O2 by a formal condition we will call *property continuity*.

O1 and O2 are conflicting objectives. Each one, individually, has a trivial solution: O1 will be satisfied by a "minimal" distance that is constant zero (we will allow distances that are not metrics, and where non-identical models can have distance zero). O2, on the other hand, is satisfied by any "maximal" distance, where any two non-identical models have the maximal possible distance, typically 1 or ∞. The challenge, then, is to find distances that in a meaningful manner balance O1 and O2.

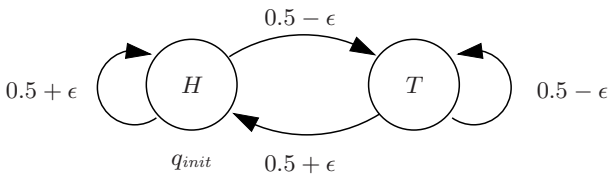

Fig. 1. Biased coin model $\mathcal{M}_\epsilon$

In this paper we investigate how a number of different distance functions perform with regard to the criteria of parameter and property continuity. Besides the existing bisimulation distances, our main interest is with *trace-based* distances that measure the distance between automata only as a function of the probability distributions over infinite sequences defined by the automata. We here study several constructions of distance measures derived from the standard L_p and Kullback-Leibler distances. It will turn out that the conflict between O1 and O2 is not fully resolved by any of our candidate distance measures, and we will derive an impossibility result that explains to some extent the fundamental difficulty we encounter.

2 Preliminaries

Throughout, Σ denotes a finite alphabet; $\Sigma^n, \Sigma^*, \Sigma^\omega$ denote the sets of all strings of length n, all finite strings, and all infinite strings, respectively. A finite string $w \in \Sigma^*$ defines the *cylinder set* $w\Sigma^\omega \subseteq \Sigma^\omega$. This is just the set of all infinite strings with prefix w. Let *Cyl* denote the set of all cylinder sets. *Cyl* is the basis of the standard topology $\mathcal{O}(\Sigma^\omega)$ on Σ^ω, i.e., open sets in this topology are just unions of cylinder sets. Furthermore, the cylinder sets generate the σ-algebra $\mathcal{A}(\Sigma^\omega)$ on Σ^ω.

The basic automaton model we use in this work is that of a *Labeled Markov Chain*, or, more specifically, state-labelled, discrete time Markov chain.

Definition 1. *LMC A* labeled Markov chain (LMC) *over Σ is a tuple $\mathcal{M} = \langle Q, \Sigma, \Pi, \pi, L\rangle$, where*

- *Q is a finite set of states,*
- *$\Pi : Q \to [0,1]$ is an* initial probability distribution *with $\sum_{q\in Q} \Pi(q) = 1$,*
- *$\pi : Q \times Q \to [0,1]$ is the* transition probability function *such that for all $q \in Q$, $\sum_{q'\in Q} \pi(q,q') = 1$.*
- *$L : Q \to \Sigma$ is a* labeling function

If $\Pi(q_{\text{init}}) = 1$ for some unique initial *state q_{init} of Q, then we denote $\mathcal{M}$ also as $\langle Q, \Sigma, q_{\text{init}}, \pi, L\rangle$. In contexts where the initial distribution Π does not matter, we also simply consider the structure $\langle Q, \Sigma, \pi, L\rangle$ as a LMC.*

An LMC is deterministic *if a state $q_{\text{init}} \in Q$ as described above exists, and for all $q \in Q$, $\sigma \in \Sigma$ there exists at most one state $q' \in Q$ with $\pi(q,q') > 0$ and $L(q') = \sigma$.*

According to the preceding definition, we assume that each state is labelled with a unique symbol from Σ, not by a subset of a set of atomic propositions AP, as, e.g., in [2]. Clearly, by taking $\Sigma = 2^{AP}$, Definition 1 also accommodates this alternative view of labeled Markov chains.

An LMC defines for each n a probability distribution over Q^n, which induces via the mapping $q_{i_1} \dots q_{i_n} \mapsto L(q_{i_1}) \dots L(q_{i_n})$ a probability distribution on Σ^n. Via standard measure-theoretic constructions, these distributions define a unique

distribution on $\mathcal{A}(\Sigma^\omega)$, which we denote by $P_{\mathcal{M}}$. When the initial distribution Π of $\mathcal{M}$ is re-defined to assign probability one to $q \in Q$, then we denote the distribution defined by the resulting LMC by $P_{\mathcal{M},q}$. This can be simplified to P_q, when the underlying structure $\langle Q, \Sigma, \cdot, \pi, L\rangle$ is clear from the context.

Linear-time properties related to traces of the model can be expressed in linear-time temporal logic (LTL) enriched also with the derived temporal operators $\Box$ (always) and $\Diamond$ (eventually). The fragment of LTL obtained by omitting the until operator $\varphi_1 \mathsf{U} \varphi_2$ is called *bounded LTL (BLTL)*.

3 From Equivalence to Distance

The most fundamental approach to comparing system models is by means of concepts of system equivalence. For non-probabilistic system models, the basic tools here are bisimulation and trace equivalence. Adapted to probabilistic system models, this gives rise to the following two notions of equivalence.

Definition 2 (Probabilistic Bisimulation [10]). *Let $\mathcal{M} = \langle Q, \Sigma, \Pi, \pi, L\rangle$ be an LMC. A probabilistic bisimulation on $\mathcal{M}$ is an equivalence relation R on Q such that for all states $(q_1, q_2) \in R$:*

- $L(q_1) = L(q_2)$.
- $\pi(q_1, C) = \pi(q_2, C)$ *for each equivalence class* $C \in Q/R$.

States q_1 and q_2 are bisimulation-equivalent (or bisimilar), denoted $q_1 \sim q_2$, if there exists a bisimulation R on $\mathcal{M}$ such that $(q_1, q_2) \in R$.

Definition 3 (Probabilistic Trace Equivalence). *Two states $q_1 \in \mathcal{M}_1, q_2 \in \mathcal{M}_2$ are probabilistic trace equivalent, denoted $q_1 \overset{T}{\sim} q_2$, if $P_{\mathcal{M}_1,q_1} = P_{\mathcal{M}_2,q_2}$.*

Equivalence often is too strong a condition when comparing system models. We therefore also need quantitative measures that allow us to determine whether one system very closely resembles another system, without being completely indistinguishable in the sense of an equivalence relation. We study such measures given in the form of distance functions, where small distance indicates similarity, and zero distance means equivalence.

Thus, we consider *distance functions* d that map pairs of states to non-negative numbers: $d : (q_1, q_2) \to \mathbb{R}^{\geq 0} \cup \{\infty\}$. The only condition we always require is that $d(q, q) = 0$. If $d < \infty$, d is symmetric and satisfies the triangle inequality, then d is called a *pseudo-metric*. If also $q_1 \neq q_2 \Rightarrow d(q_1, q_2) > 0$, then d is a *metric*. We note that as a measure of approximation quality, non-symmetric distances can be quite natural, because here the two arguments of the distance function can have distinct roles: one being the approximation, and one being the "real" model that is approximated. For example, if ϕ expresses a crucial safety property, and $\mathcal{M}_1, \mathcal{M}_2$ are LMCs with $P_{\mathcal{M}_1}(\neg\phi) = 0$ and $P_{\mathcal{M}_2}(\neg\phi) = 10^{-5}$, then $\mathcal{M}_2$ may be considered a good (i.e., safe) approximation of $\mathcal{M}_1$, but not vice-versa.

A distance function d is *consistent with bisimilarity* if $d(q_1, q_2) = 0 \Leftrightarrow q_1 \sim q_2$; it is *consistent with trace equivalence* if $d(q_1, q_2) = 0 \Leftrightarrow q_1 \overset{T}{\sim} q_2$. If a distance is consistent with trace equivalence, then the implication $q_1 \sim q_2 \Rightarrow d(q_1, q_2) = 0$ still holds, but not the converse.

We next introduce two types of distance functions. First we consider distance functions that are quantitative extensions of bisimulation equivalence, and then distance functions that extend trace equivalence.

3.1 Probabilistic Bisimilarity Metric

The bisimilarity pseudometric was originally introduced by means of logical expressions that are evaluated to real numbers at system states according to a functional semantics [6, 20]. The distance of two states then is defined as the supremum over all logical expressions of the differences of function values. Alternative characterizations as the fixedpoint of monotone operators on pseudometrics have been developed in [19, 3, 1].

However, there are some differences in the assumed underlying system models in these papers, and the literature does not fully establish an equivalence of all available versions of bisimilarity distance for the LMC models we here use. In the following we therefore review one particular formalization of the bisimilarity pseudometrics in terms of *couplings*.

The definitions below follow the style traditionally used in the bisimulation context in that a distance is defined between states q_1, q_2 in a single underlying model $\mathcal{M}$. There is just a minor conceptual difference with no technical implications between this perspective, and the view that each q_i is embedded in its own model $\mathcal{M}_i$.

Given two probability measures μ, ν on Q, we use the notation $J(\mu, \nu)$ to denote the set of all probability measures on $Q \times Q$ that have μ and ν as the marginals on the first, respectively second, component.

Definition 4 (Coupling). *Let $\mathcal{M} = \langle Q, \Sigma, \pi, L\rangle$ be a finite LMC. The Markov chain $\mathcal{C} = \langle Q \times Q, \Sigma \times \Sigma, \omega, L\rangle$ is called a* coupling for $\mathcal{M}$ *if, for all $q_1, q_2 \in Q$,*

1. $\omega((q_1, q_2), \cdot) \in J(\pi(q_1, \cdot), \pi(q_2, \cdot))$, *and*
2. $L(q_1, q_2) = (L(q_1), L(q_2))$.

A coupling for $\mathcal{M}$ can be seen as a probabilistic pairing of two copies of $\mathcal{M}$ running synchronously, although not necessarily independently.

Given a coupling $\mathcal{C}$ for $\mathcal{M}$, and a *discount factor* $\lambda \leq 1$, we define $\Gamma^{\mathcal{C}}_{\lambda} \colon [0,1]^{Q\times Q} \to [0,1]^{Q\times Q}$ for $d \colon Q \times Q \to [0,1]$ and $q_1, q_2 \in Q$, as follows:

$$\Gamma^{\mathcal{C}}_{\lambda}(d)(q_1, q_2) = \begin{cases} 1 & \text{if } L(q_1) \neq L(q_2) \\ \lambda \cdot \sum_{u,v \in Q} d(u, v) \cdot \omega((q_1, q_2), (u, v)) & \text{if } L(q_1) = L(q_2) \end{cases}$$

The operator $\Gamma^{\mathcal{C}}_{\lambda}$ has a unique least fixedpoint [1], which we denote by $\gamma^{\mathcal{C}}_{\lambda}$. Each $\gamma^{\mathcal{C}}_{\lambda}$ is a distance function on Q. The bisimulation distance is obtained by

taking the minimum over all possible couplings:

$$d_{b,\lambda} := \min\{\gamma_\lambda^{\mathcal{C}} \mid \mathcal{C} \text{ coupling for } \mathcal{M}\}. \tag{1}$$

The minimum here is taken pointwise at each argument (q_1, q_2). It is shown in [1] that $d_{b,\lambda}$ is well-defined, as the minimum on the right of (1) is attained. Furthermore, there is a coupling that minimizes $\gamma_\lambda^{\mathcal{C}}(q_1, q_2)$ simultaneously for all (q_1, q_2). We here use the extra subscript b to distinguish this bisimilarity distance more clearly from other distance functions we will also consider in the sequel. $d_{b,\lambda}$ is consistent with probabilistic bisimilarity.

3.2 Trace-based Distances

A distance d is trace-based, if $d(q_1, q_2)$ is a function only of P_{q_1} and P_{q_2}. The measure-theoretic construction of distributions P_q on Σ^ω is essentially a limit of finite-dimensional distributions on Σ^n ($n \in \mathbb{N}$). In a similar manner, it is natural to construct distances between distributions on Σ^ω as a limit of distances on distributions on Σ^n. There are, however, several possible ways of doing this. We consider the following three canonical constructions. If $d^{(n)}$ is a distance function for distributions on Σ^n ($n \in \mathbb{N}$), we define induced distance functions for distributions on Σ^ω as

- (limit) $d^\infty := \lim_{n\to\infty} d^{(n)}$
- (per-symbol distance; limit average) $d^{ps} := \lim_n \frac{1}{n} d^{(n)}$
- (discounted sum) $d^\lambda := \sum_{n\geq 1} \lambda^n d^{(n)}$ $(\lambda < 1)$

For all three constructions it holds that symmetry and triangle inequality are preserved, i.e., if all the $d^{(n)}$ possess these properties, then so do d^λ, d^∞, and d^{ps} (provided the limits exist).

The limit and the per-symbol distances are opposite in nature to the discounted sum distances: the latter emphasizes the differences in the distribution of initial segments $w \in \Sigma^n$ of $s = ws' \in \Sigma^\omega$, whereas the first two are most sensitive to the distribution of the infinite tail s'.

In the following P_1, P_2 always denote probability distributions on Σ^ω. We are mostly concerned with distributions P_i that are defined by states q_i in LMCs $\mathcal{M}_i$, i.e., $P_i = P_{q_i}$. However, many of our general considerations also apply to arbitrary distributions P_i. If P_i is of the form P_{q_i} for some $q_i \in \mathcal{M}_i$, then we say that P_i is generated by an LMC.

P_i induces for each $n \in \mathbb{N}$ a distribution $P_i^{(n)}$ on Σ^n. In order to avoid notational clutter, we suppress the superscript (n) to distinguish $P_i^{(n)}$ from P_i. Which probability space we assume for P_i in a given context will be implicit from the arguments of $P_i()$.

In this paper, we consider the following standard distance functions between distributions P_1, P_2 on Σ^n:

- (Kullback-Leibler distance) $d_{KL}^{(n)}(P_1, P_2) := \sum_{w\in\Sigma^n} P_1(w) \log \frac{P_1(w)}{P_2(w)}$

– (L_p-distance) $d^{(n)}_{L_p}(q_1, q_2) := (\sum_{w\in\Sigma^n} \mid P_1(w) - P_2(w) \mid^p)^{1/p}$

For L_p-distances we focus our attention on $p = 1$ (total variation distance), $p = 2$ (Euclidean distance), and $p = \infty$ (Maximum distance). The distance d^{ps}_{KL} is well-known in information theory, and there usually called the *divergence-rate*.

An important tool in the analysis of the Kullback-Leibler distance is an additivity property [9, Chapter 2], which adapted to our context can be stated as:

$$d^{(n+1)}_{KL}(P_1, P_2) = d^{(n)}_{KL}(P_1, P_2) + \sum_{w\in\Sigma^n} P_1(w) \sum_{\sigma\in\Sigma} P_1(\sigma|w) \log \frac{P_1(\sigma|w)}{P_2(\sigma|w)}. \tag{2}$$

Also important is the following relationship between d_{KL} and d_{L_1}:

$$d^{(n)}_{KL} \geq \frac{(d^{(n)}_{L_1})^2}{2}. \tag{3}$$

(see [18] for this and further, sharper, bounds). A direct implication is that for all $A \subseteq \Sigma^n$

$$|\, P_1(A) - P_2(A) \,| \leq \sqrt{d^{(n)}_{KL}(P_1, P_2)/2}. \tag{4}$$

For all definitions of distances as limits one needs to verify that the limits actually exist in order to ensure that the distances are well-defined. The following table summarizes some relevant facts:

	KL	L_1	L_2	L_∞
d^λ	Lemma 1	$< \infty$	$< \infty$	$< \infty$
d^∞	Lemma 1	$< \infty$	?	?
d^{ps}	Prop. 1	$\equiv 0$	$\equiv 0$	$\equiv 0$

Here '$< \infty$' means that the distance is well defined and finite. For the $d^\lambda_{L_p}$ distances this is the case because the $d^{(n)}_{L_p}$ are bounded by a common constant for all n. For L_1 one furthermore has that $d^{(n)}_{L_1}$ is monotonically increasing in n, which entails $d^\infty_{L_1} < \infty$. $d^{(n)}_{L_2}$ and $d^{(n)}_{L_\infty}$ are not monotone in n. From Proposition 2 below it follows that $d^\infty_{L_2}, d^\infty_{L_\infty}$ will be not very useful even if guaranteed to be well-defined. We therefore do not analyse their exact status further.

The $d^{(n)}_{L_p}$ being bounded, it is also immediate that the $d^{ps}_{L_p}$ are identically zero, denoted $\equiv 0$ in the table. We will not consider these distances further.

We now turn to the limiting behavior of $d^{(n)}_{KL}$, where the situation is a little more intricate.

Lemma 1. **(i)** *$d^{(n)}_{\mathrm{KL}}(P_1, P_2)$ is monotonically increasing for all P_1, P_2.*

If P_1, P_2 are generated by LMCs, then one of the following cases holds:

(iia) *There exists an $n > 0$ and $w \in \Sigma^n$ with $0 = P_2(w) < P_1(w)$, so that $d^{(m)}_{\mathrm{KL}}(P_1, P_2) = \infty$ for all $m \geq n$.*

(iib) $d^{(n)}_{\mathrm{KL}}(P_1, P_2) \in O(n)$

From this Lemma it follows that d^{λ}_{KL} and d^{∞}_{KL} are well-defined, but possibly infinite.[3] Furthermore, if case (iib) holds, then d^{λ}_{KL} is finite.

Turning to the per-symbol distance, we first obtain from Lemma 1 that if the P_i are generated by LMCs, and case (iia) of the lemma does not hold, then

$$0 \leq \liminf_n \frac{1}{n} d^{(n)}_{KL}(P_1, P_2) \leq \limsup_n \frac{1}{n} d^{(n)}_{KL}(P_1, P_2) < \infty.$$

To ensure that d^{ps}_{KL} is well-defined, one has to establish that the lim inf and lim sup are equal in this equation. The question of this equality, i.e., the problem of the existence of the divergence rate, is non-trivial, and has received considerable attention in the literature. [15] gives examples of stochastic processes for which the divergence rate does not exist, but also states that it exists when P_1, P_2 are generated by Hidden Markov Models (HMMs). Since LMCs are a special type of Hidden Markov Models, this would provide the solution to our problem. However, no proof of this statement is given in [15]. Positive results on the existence of the divergence rate for several classes of Markov processes can be found in [13] and [8, Chapter 10]. These results do not cover the case of HMMs or LMCs, however. In contrast, [7, 16] specifically consider the class of HMMs, but the results of [7] applied to our problem will only lead to the trivial bound $\limsup_n \frac{1}{n} d^{(n)}_{KL}(P_1, P_2) \leq \infty$, and [16] is concerned with models with continuous observation spaces.

We will not solve the question of the existence of d^{ps}_{KL} in full generality here. In the following, we only consider the case of deterministic LMCs. This case not only greatly facilitates the theoretical analysis, but the proof of the following Proposition also leads to an efficient way of computing d^{ps}_{KL}. [4]

Proposition 1. *Let P_1, P_2 be defined by deterministic LMCs $\mathcal{M}_1, \mathcal{M}_2$. Then $\lim_n 1/n \; d^{(n)}_{\mathrm{KL}}(P_1, P_2)$ exists.*

In light of [15] it is strongly conjectured that the existence of $d^{ps}_{KL}(P_1, P_2)$ also holds for nondeterministic LMCs. In the following, all statements relating to d^{ps}_{KL} are implicitly restricted to those cases where d^{ps}_{KL} is well-defined.

Having defined several candidate trace-based distances, we first check which ones are consistent with trace equivalence.

Proposition 2. *Distances are or are not consistent with trace equivalence, as indicated by y (yes), respectively n (no), in the following table:*

[3] The proof of this lemma and subsequent results can be found in the online appendices for this paper available at `people.cs.aau.dk/~jaeger/publications.html`

[4] We note that the efficient computability of the finite-dimensional $d^{(n)}_{KL}$, as well as the limits d^{∞}_{KL} and d^{ps}_{KL} is a different problem than the computation of relative entropies for probabilistic automata, as investigated by [5]. The automata investigated in this latter work define probability distributions over Σ^*, and the Kullback-Leibler distance therefore becomes the infinite sum $\sum_{w \in \Sigma^*} P_1(w) \log(P_1(w)/P_2(w))$.

	KL	L_1	L_2	L_∞
d^λ	y	y	y	y
d^∞	y	y	n	n
d^{ps}	n			

For $d^\infty_{L_2}$ and $d^\infty_{L_\infty}$ the proposition is shown by considering the automata of Figure 1: denote by q_ϵ the initial state of automaton $\mathcal{M}_\epsilon$. Then one obtains that for all ϵ $d^\infty_{L_2}(\mathcal{M}_0, \mathcal{M}_\epsilon) = d^\infty_{L_\infty}(\mathcal{M}_0, \mathcal{M}_\epsilon) = 0$. Not being able to measure any distance between different $\mathcal{M}_\epsilon$ models makes these distance measure clearly unsuitable for our purpose, and we will not consider them any further.

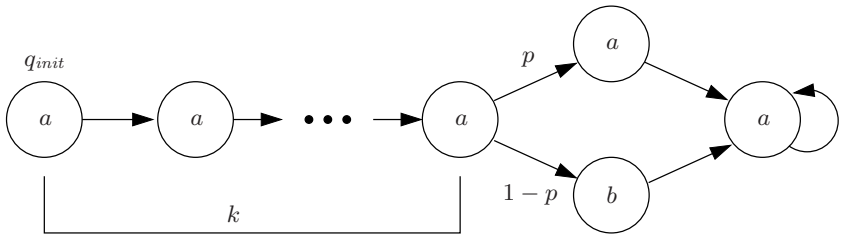

Fig. 2. The automata $\mathcal{M}_{k,p}$

According to Proposition 2, also d^{ps}_{KL} is not consistent with trace equivalence. An example illustrating this point is given by Figure 2. It shows a (deterministic) LMC $\mathcal{M}_{k,p}$ parameterized by k (length of an initial sequence of a-labeled states), and p (the indicated transition probability). Consider the case $k = 1$. Let $p \neq p'$, and q, q' the initial states of $\mathcal{M}_{1,p}$ and $\mathcal{M}_{1,p'}$, respectively. Then one obtains that $d^{(n)}_{KL}(q, q') = p \log(p/p') + (1-p)\log((1-p)/(1-p'))$ is constant for all n, so that $d^{ps}_{KL}(q, q') = 0$.

Even though d^{ps}_{KL} here also fails to distinguish different models $\mathcal{M}_{k,p}$ and $\mathcal{M}_{k,p'}$, this failure is much less significant than the failure of $d^\infty_{L_2}$ and $d^\infty_{L_\infty}$ for the models $\mathcal{M}_\epsilon$. The $\mathcal{M}_{k,p}$ are indeed only distinguishable by their initial behavior, but indistinguishable in their infinitary, ergodic behavior. If one is primarily concerned with the limiting behavior of systems, then $d(q, q') = 0$ is appropriate. For $\mathcal{M}_0$ and $\mathcal{M}_\epsilon$ of Figure 1 the ergodic behaviors are characterized by a different frequency of H and T, and $d^{ps}_{KL}(q_0, q_\epsilon) = 0.5 \log(0.5/(0.5+\epsilon)) + 0.5 \log(0.5/(0.5-\epsilon))$ appropriately reflects this.

We therefore still consider d^{ps}_{KL} as a meaningful distance. Even if we insist on consistency with trace equivalence as a necessary property for a distance, d^{ps}_{KL} remains relevant for the following reason: if d is a distance that is consistent with trace equivalence, then any mixture $\alpha d + (1-\alpha)d'$ $(0 < \alpha < 1)$ of d with another distance d' still is consistent with trace equivalence. Thus, even if d^{ps}_{KL} may not satisfy our demands for a stand-alone distance, it can still be a very useful component in a distance defined as a mixture. We will return to the construction of distances as mixtures in Section 6.

We here have considered constructions of distance functions for distributions on Σ^ω from distance functions on Σ^n. Of course, one may also directly define

distances on Σ^ω using integrals rather than sums. For example, one may define

$$d_{KL}(P_1, P_2) = \int_{\Sigma^\omega} f_1(s) \log \frac{f_1(s)}{f_2(s)} d\mu(s),$$

where the f_i are density functions for P_i relative to the reference measure μ. For this, however, according to the Radon-Nikodym theorem, we first need a reference measure μ, so that the P_{q_i} are both absolutely continuous with respect to μ. In general, it will be impossible to find a natural μ that serves this purpose for all relevant P_i. However, one can work around this problem by letting $\mu = 1/2(P_1 + P_2)$. Distances defined in this manner, however, will fail our first desirable property, introduced in the following section.

4 Main Properties

4.1 Parameter Continuity

We begin by giving a general formalization of the intuition that as $\epsilon \to 0$ in Figure 1, the distance between the corresponding states of $\mathcal{M}_0$ and $\mathcal{M}_\epsilon$ should go to zero.

Let π be a transition probability function on a state set Q. A sequence $(\pi_n)_n$ of transition probability functions on Q *s-converges* against π, denoted $\pi_n \xrightarrow{s} \pi$, if

(i) $\forall n \forall q, q' \in Q : \ \pi_n(q, q') = 0 \Leftrightarrow \pi(q, q') = 0$
(ii) $\forall q, q' \in Q : \ \pi_n(q, q') \to \pi(q, q') \ (n \to \infty)$

We call this s-convergence, because condition (i) requires that the functions in the sequence (π_n) all have the same set of support as π. In other words, we do not allow a sequence of non-zero transition probabilities to converge to zero.

Definition 5. *A distance function d is* parameter continuous, *if for any labeled Markov chain $\mathcal{M} = \langle Q, \Sigma, q_{\text{init}}, \pi, L\rangle$, and any sequence $\pi_n \xrightarrow{s} \pi$ the following holds: for $\mathcal{M}_n := \langle Q, \Sigma, q_{\text{init}}, \pi_n, L\rangle$, $P := P_{\mathcal{M}, q_{\text{init}}}$, and $P_n := P_{\mathcal{M}_n, q_{\text{init}}}$ it holds that $\lim_{n\to\infty} d(P_n, P) = \lim_{n\to\infty} d(P, P_n) = 0$.*

Note that we are considering potentially non-symmetric distance functions, which is why we have the requirements both for the limit of $d(P_n, P)$ and $d(P, P_n)$.

Parameter continuity in the sense of this definition captures an important aspect of the informal objective O1 from Section 1. We only consider s-convergent sequences of transition probabilities in this definition, because a stronger requirement that also applies to sequences of transition probabilities $\epsilon_n \to 0$ would be immediately inconsistent with objective O2, as formalized by LTL-continuity below: consider the coin model of Figure 1, but now let the transition probabilities into the H state be $1 - \epsilon$, and the transition probabilities into T be ϵ. For the LTL property $\Diamond T$ we then have $P(\Diamond T) = 1$ in all $\mathcal{M}_\epsilon$ with $\epsilon > 0$, and

$P(\Diamond T) = 0$ in $\mathcal{M}_0$. Thus, if we required that $d(\mathcal{M}_\epsilon, \mathcal{M}_0) \to 0$ as $\epsilon \to 0$, then an upper bound on the distance between models could not imply an upper bound on the probability difference for LTL formulas.

The following Proposition summarizes parameter continuity properties of selected distances. We do not consider any more those trace-based distances that from Proposition 2 turned out to be uninteresting.

Proposition 3. *Distances are or are not parameter continuous, as indicated by y (yes), respectively n (no), in the following tables:*

	KL	L_1	L_2	L_∞
d^λ	y	y	y	y
d^∞	n	n		
d^{ps}	y			

	$d_{b,\lambda}$
$\lambda = 1$	n
$\lambda < 1$	y

The negative result for $d_{b,1}$ is obtained from a characterization of $d_{b,1}$ in terms of the reachability probability in couplings $\mathcal{C}$ of a state (q_1, q_2) with $L(q_1) \neq L(q_2)$ [3, 1]. Applied to the models $\mathcal{M}_\epsilon$ of Figure 1, this characterization shows that $d_{b,1}(q_0, q_\epsilon) = 1$ for all $\epsilon > 0$.

The negative results for d^∞_{KL} and $d^\infty_{L_1}$ are also obtained by considering the models $\mathcal{M}_\epsilon$, where one again obtains that distances between $\mathcal{M}_0$ and $\mathcal{M}_\epsilon$ are given by the maximal possible values: for all $\epsilon > 0$ $d^\infty_{KL}(q_0, q_\epsilon) = \infty$, $d^\infty_{L_1}(q_0, q_\epsilon) = 2$.

4.2 Property Continuity

In the following, we call any measurable subset $\varphi \subseteq \Sigma^\omega$ a *property*. Thus, "property" is the same as "event" in standard probability theoretic language. We prefer the term property here, because in the present context we view φ rather as a property of a system behavior than as an observed event, and it will later be more natural to speak about LTL-definable properties, than LTL-definable events.

Definition 6 (Φ-continuity). *Let $\varphi \subseteq \Sigma^\omega$ be a property. A distance d is* φ-continuous*, if*

$$\forall \epsilon > 0\ \exists \delta > 0\ \forall P_1, P_2 : d(P_1, P_2) \leq \delta \Rightarrow\ |P_1(\varphi) - P_2(\varphi)| \leq \epsilon. \tag{5}$$

If $\Phi \subset 2^{\Sigma^\omega}$ is a class of properties, then d is Φ-continuous*, if d is φ-continuous for all $\varphi \in \Phi$.*

If d is *Φ-continuous*, then the δ-bound on $d(P_1, P_2)$ needed to ensure that $|P_1(\varphi) - P_2(\varphi)| \leq \epsilon$ will depend on φ. In the following definition these bounds are required to be uniform for all φ.

Definition 7 (Uniform Φ-continuity). *Let $\Phi \subset 2^{\Sigma^\omega}$ be a class of properties. A distance d is* uniformly *Φ-continuous, if*

$$\forall \epsilon > 0\ \exists \delta > 0\ \forall \varphi \in \Phi, \forall P_1, P_2 : d(P_1, P_2) \leq \delta \Rightarrow\ |P_1(\varphi) - P_2(\varphi)| \leq \epsilon. \tag{6}$$

The following lemma is a straightforward, but useful observation.

Lemma 2. *Let d_1, d_2 be two distance function, such that there exists a continuous function f with $f(0) = 0$, and $d_1 \leq f(d_2)$. Then, for any Φ: (uniform) Φ-continuity of d_1 implies (uniform) Φ-continuity of d_2.*

According to (3) Lemma 2 applies to $d_1 = d^{(n)}_{L_1}$ and $d_2 = d^{(n)}_{KL}$ with $f(x) = \sqrt{2x}$. Since f does not depend on n, the same also is true for $d_1 = d^{\infty}_{L_1}$ and $d_2 = d^{\infty}_{KL}$. Thus, proving (uniform) Φ-continuity for $d^{\infty}_{L_1}$ is sufficient to also prove it for d^{∞}_{KL}.

Lemma 3. *A BLTL-definable property $\phi \subseteq \Sigma^{\omega}$ is a finite union of cylinder sets. A distance d is (uniformly) BLTL continuous iff it is (uniformly)* Cyl*-continuous.*

The first statement in this lemma follows from a straightforward induction on BLTL formulas. The second statement then is a direct consequence of the definitions. Combining Lemma 3 with the fact that measures on Σ^{ω} are uniquely defined by the measures of cylinder sets, one obtains:

Lemma 4. *BLTL-continuity implies consistency with trace equivalence.*

We now formulate our main results on property continuity.

Proposition 4. *Distances are or are not uniformly BLTL continuous, BLTL continuous, or not BLTL continuous as indicated by uy, y, respectively n, in the following tables:*

	KL	L_1	L_2	L_∞
d^{λ}	y	y	y	y
d^{∞}	uy	uy		
d^{ps}	n			

	$d_{b,\lambda}$
$\lambda = 1$	uy
$\lambda < 1$	y

Proposition 5. *Distances are or are not uniformly LTL continuous, LTL continuous, or not LTL continuous, as indicated by uy, y, respectively n, in the following tables:*

	KL	L_1	L_2	L_∞
d^{λ}	n	n	n	n
d^{∞}	uy	uy		
d^{ps}	n			

	$d_{b,\lambda}$
$\lambda = 1$	uy
$\lambda < 1$	n

The negative results for the d^{λ} distances are established by again considering the automata $\mathcal{M}_{k,p}$ of Figure 2, and the LTL sentence $\Diamond b$. Let $p_1 \neq p_2$, and $P_{k,i}$ the distribution defined by $\mathcal{M}_{k,p_i}$ $(i = 1, 2)$ Then, for all k: $\mid P_{k,1}(\Diamond b) - P_{k,2}(\Diamond b) \mid = \mid p_1 - p_2 \mid$. On the other hand, for all discounted distances, and all $\delta > 0$, there exists a k such that $d(P_{k,1}, P_{k,2}) < \delta$.

According to Proposition 5, $d^{\infty}_{KL}, d^{\infty}_{L_1}$ and $d_{b,1}$ have very strong property continuity characteristics. However, according to Proposition 3, this comes at the price of not fulfilling objective O1.

Comparison of Propositions 3 and 5 shows that so far we have failed to construct a distance function implementing both our objectives. In the following section we will see that to some extent this is due to a fundamental limitation.

5 Impossibility Results

The proofs of the positive results expressed by Propositions 4 and 5 are not based on the concrete logical characterizations of property classes LTL and BLTL, but on the underlying topological and measure-theoretic structure of these properties. This is not surprising, since the definitions of the distance measures we have been considering also are based on general measure-theoretic concepts, without reference to linear temporal logic.

It is therefore tempting to try to construct on a slightly broader topological basis also a distance measure that is both parameter- and LTL-continuous. This approach also suggests itself because of the fact that LTL-definable properties still have a quite simple topological structure: any LTL-definable set $A \subseteq \Sigma^\omega$ is a Boolean combination of G_δ-sets, where a G_δ-set is a countable intersection of open sets [17, Theorem 5.2]. From this it follows that for LTL-continuity of d it would be enough to show that d is continuous for G_δ-sets.

However, as we now show, it is even impossible to obtain continuity for all open sets in conjunction with parameter continuity.

Proposition 6. *There exists an open set O, such that there does not exist a distance function d that is parameter continuous and O-continuous.*

The open set O constructed in the proof of the preceding theorem is not LTL-definable. The theorem, therefore, only delimits the possibilities of obtaining parameter continuous and LTL-continuous distance functions by purely topological and measure-theoretic constructions. However, the proof of Proposition 6 also directly leads to the following:

Proposition 7. *There does not exist a distance function d that is parameter continuous and uniformly BLTL-continuous.*

Thus, we find that uniform (B)LTL-continuity is inconsistent with parameter continuity. However, uniform continuity is a very strong demand to begin with, so the main objective of combining parameter continuity and LTL-continuity still could be feasible. In the next section we show that this is indeed the case. Before we give a concrete example, we establish some general results about mixtures of distance functions.

6 Mixture Constructions

In Section 3 we justified our continued interest in the d_{KL}^{ps} distance in spite of the fact that it is not consistent with trace equivalence by its possible use as a component in a mixture of distances.

Definition 8. *Let $n \in \mathbb{N} \cup \{\infty\}$, and for $i = 1, \ldots, n$: $\alpha_i \in [0,1]$ with $\sum_i \alpha_i = 1$, and d_i a distance function. Then $d := \sum_i \alpha_i d_i$ is called a* mixture *of the d_i. If $n < \infty$, then d is a* finite mixture.

It is well-known that mixtures of distances preserve essential metric properties such as symmetry and the triangle-inequality. In the following we summarize to what extent the distance properties we are studying are preserved. We say that a property of a distance is *preserved* under mixtures, if a mixture d has the property whenever all its constituent d_i have the property. A property is *strongly preserved* if d has the property whenever at least one d_i has the property. The following Lemma summarizes the relevant preservation properties.

Lemma 5. *The following properties are preserved under mixtures:*

- *The left to right direction* $d(P_1, P_2) = 0 \Leftarrow q_1 \equiv q_2$ *(*$\equiv \in \{\sim, \overset{T}{\sim}\}$*) of consistency with bisimulation- or trace-equivalence.*
- *Parameter continuity*

The following properties are strongly preserved under mixtures:

- *The right to left direction* $d(P_1, P_2) = 0 \Rightarrow q_1 \equiv q_2$ *(*$\equiv \in \{\sim, \overset{T}{\sim}\}$*) of consistency with bisimulation- or trace-equivalence.*
- *Φ-continuity and uniform Φ-continuity.*

We next investigate two different distances that are constructed as mixtures.

6.1 Expected LTL Distance

Definition 9. *For* $\phi \in$ LTL *define*

$$d_\phi(P_1, P_2) := |P_1(\phi) - P_2(\phi)|.$$

Let $\phi_1, \phi_2, \ldots$ *be an enumeration of LTL,* $\alpha_i \in (0, 1)$ *with* $\sum_i \alpha_i = 1$*, and define*

$$d_Q := \sum_i \alpha_i d_{\phi_i}.$$

$d_Q(P_1, P_2)$ can be interpreted as the expected difference $|P_1(\phi) - P_2(\phi)|$ for LTL formulas that are randomly generated according to probabilities α_i. As an empirical evaluation measure for how well a learned system model approximates the LTL properties of a true system model $\mathcal{M}_1$, this distance was used in [11]. The following proposition now states that with d_Q we have the first distance that satisfies both of our main objectives.

Proposition 8. *d_Q is parameter and LTL continuous.*

Even though d_Q satisfies our main objectives, it clearly still has some significant shortcomings. First, the concrete values of $d_Q(P_1, P_2)$ depend very much on the coefficients α_i. When the α_i are just more or less arbitrarily set in a synthetic construction of d_Q, then the actual values of d_Q will lack a meaningful interpretation. If, however, α_i represents a meaningful probability of ϕ_i (for example, the expected frequency with which ϕ_i will be checked in a given

application context), then $d_Q(P_1, P_2)$ is interpretable as the expected deviation between LTL-probabilities computed in $\mathcal{M}_1$ and $\mathcal{M}_2$.

Second, d_Q poses computational problems. The only currently available approach to (approximately) computing d_Q is to compute d_{ϕ_i} for a sample $i = i_1, \ldots, i_k$. If α_i can be computed for a given ϕ_i, and the ϕ_{i_j} in the sample are all distinct, then d_Q is bounded by $[\sum_{j=1}^{k} \alpha_{i_j} d_{\phi_{i_j}}, \sum_{j=1}^{k} \alpha_{i_j} d_{\phi_{i_j}} + (1 - \sum_{j=1}^{k} \alpha_{i_j})]$. If the α_i are only implicitly given by a random generator for LTL formulas, then d_Q can be estimated by the empirical distance $1/k \sum_{j=1}^{k} d_{\phi_{i_j}}$.

6.2 KL mixture

A second mixture construction we consider is

$$d_{KL}^{mix} := \alpha d_{KL}^{\lambda} + (1 - \alpha) d_{KL}^{ps}.$$

The motivation for d_{KL}^{mix} is that it combines a distance function that is mostly sensitive to differences in the initial behavior of a system (d_{KL}^{λ}), and a distance that measures differences in the long-run, ergodic behavior (d_{KL}^{ps}).

d_{KL}^{mix} is consistent with trace-equivalence, parameter continuous, and inherits the BLTL-continuity of d_{KL}^{λ}. However, d_{KL}^{mix} is not LTL continuous (as expected from Proposition 6, since d_{KL}^{mix} is a purely measure theoretic construction). Concretely, d_{KL}^{mix} still is subject to the counterexample described for the d^{λ} in connection with Proposition 5.

7 Conclusion

In this paper we have investigated a number of distances on finite state Markov Processes, which measure the behavioural similarity of non-bisimilar processes. We have considered both bisimulation distances and trace-based distances. In particular, we focused on several constructions derived from the standard L_p and Kullback-Leibler distances. The continuity aspects for which we have tested the distances are natural properties one would expect from a distance that in a meaningful sense measures the relationship between a true model and its approximations. On one hand we study the parameter continuity, which guarantees that the distances are continuous in the transition probabilities. On the other hand we analyzed the concept of a good approximation of a system in the light of a given distance function. We expect from a good distance to provide us some bounds on the error incurred by using the approximation of a model instead of the real model in given contexts.

We demonstrated that none of the considered distances fully respects the continuity properties that we considered. This failure is partially explained by an impossibility result that reveals to some extent the fundamental difficulties that one encounters when trying to achieve such complex goals.

A Appendix

Proof (Proof of Lemma 1). (i) directly follows from (2). (iia) follows directly from the definition of $d_{KL}^{(n)}$ (the assumption that the P_i are defined by an LMC is not needed here). Now assume that the P_i are generated by LMCs $\mathcal{M}_i$, and that (iia) does not hold. Let $\pi_{1,max}$ be the maximal transition probability in $\mathcal{M}_1$, and $\pi_{2,min}$ the minimal non-zero transition probability in $\mathcal{M}_2$. Then

$$d_{KL}^{(n)}(P_1,P_2) = \sum_{w\in\Sigma^n:P_1(w)>0} P_1(w)\log\frac{P_1(w)}{P_2(w)} \leq \sum_{w\in\Sigma^n:P_1(w)>0} P_1(w)\log\frac{\pi_{1,max}^n}{\pi_{2,min}^n} = n\log\frac{\pi_{1,max}}{\pi_{2,min}}. \tag{7}$$

Proof (Proof of Proposition 1). First observe that the proposition is trivially true if case (iia) of Lemma 1 holds. Assume, then, that (iia) does not apply.

Let $\mathcal{M}_i = (Q_i, \Sigma, q_{init,i}, \pi_i, L_i)$ $(i = 1, 2)$. We construct a new LMC $\mathcal{M}_{1,2}$ whose states represent transitions in synchronous runs of the $\mathcal{M}_i$. More precisely, states in $\mathcal{M}_{1,2}$ are of the form (q_1, q_2, σ) where q_1, q_2 are states that are reached by a sequence of synchronized transitions, and σ is the symbol defining the next transition:

$$Q_{1,2} = \{(q_1,q_2,\sigma) \in Q_1\times Q_2\times\Sigma \mid \exists w\in\Sigma^* : q_{init,i} \overset{w}{\to} q_i \ (i=1,2); \pi_1(q_1,\sigma)>0\}$$

$$\Pi_{1,2}((q_1,q_2,\sigma)) = \begin{cases} \pi_1(q_1,\sigma) \text{ if } q_i = q_{init,i} \ (i=1,2) \\ 0 \qquad\quad \text{otherwise} \end{cases}$$

$$\pi_{1,2}((q_1,q_2,\sigma),(q_1',q_2',\sigma')) = \begin{cases} \pi_1(q_1',\sigma') \text{ if } q_1' = \pi_1(q_1,\sigma) \text{ and } q_2' = \pi_2(q_2,\sigma) \\ 0 \qquad\quad \text{otherwise} \end{cases}$$

$$L_{1,2}((q_1,q_2,\sigma)) = L_1(q_1)$$

Observe that $Q_{1,2}$ is deterministic, except for the random choice of the initial state. The transition probabilities in $Q_{1,2}$ are determined only by the transition probabilities in $\mathcal{M}_1$. Thus, $Q_{1,2}$ is essentially an augmented version of $\mathcal{M}_1$ that also tracks the state of $\mathcal{M}_2$ in a synchronized run.

We define on the states of $\mathcal{M}_{1,2}$ the function

$$f((q_1,q_2,\sigma)) := \log\frac{\pi_1(q_1,\sigma)}{\pi_2(q_2,\sigma)} \tag{8}$$

From the assumption that Lemma 1 (iia) does not hold, it follows that $f((q_1,q_2,\sigma)) < \infty$ for all states (q_1,q_2,σ). Iteratively applying (2) to decompose $d_{KL}^{(n)}$, we obtain:

$$d_{KL}^{(n)}(P_1,P_2) = E^{(n-1)}[\sum_{k=1}^{n-1} f((q_{1,k},q_{2,k},\sigma_k))]$$

where $E^{(n-1)}$ denotes the expectation in $\mathcal{M}_{1,2}$ over state sequences $((q_{1,k},q_{2,k},\sigma_k))_k$ of length $n-1$. Note that it must be the case that $L_1(q_{init,1}) = L_2(q_{init,2})$; hence

$d^{(1)}_{KL}(P_1, P_2) = 0$, and the first symbol never contributes to $d^{(n)}_{KL}(P_1, P_2)$. With $f((q_1, q_2, \sigma))$ in $\mathcal{M}_{1,2}$ we directly start measuring the KL-distance due to possible discrepancies in the label probabilities at the second states of the $\mathcal{M}_i$, which is why in (8) we have n on the left, and $n-1$ on the right side.

Let $R_1, \ldots, R_K$ be the recurrent classes of $\mathcal{M}_{1,2}$. Conditional on reaching class R_j, one has by the ergodic theorem for Markov chains (e.g. [12, Theorem 1.10.2]) that $1/n \sum_{k=1}^{n-1} f((q_{1,k}, q_{2,k}, \sigma_k))$ converges almost surely to

$$\bar{f}_k := \sum_{(q_1,q_2,\sigma)\in R_k} \rho_k((q_1, q_2, \sigma)) f((q_1, q_2, \sigma)),$$

where ρ_k is the stationary distribution on R_k. In particular, this implies that conditional on the chain reaching R_k:

$$\lim_n \frac{1}{n-1} E^{(n-1)}[\sum_{k=1}^{n-1} f((q_{1,k}, q_{2,k}, \sigma_k)) \mid R_k] = \bar{f}_k. \tag{9}$$

Thus, if α_k is the absorption probability of R_k, we obtain

$$\lim_n \frac{1}{n-1} E^{(n-1)}[\sum_{k=1}^{n-1} f((q_{1,k}, q_{2,k}, \sigma_k))] = \sum_{k=1}^{K} \alpha_k \bar{f}_k.$$

This shows that $\lim_n 1/n\, d^{(n)}_{KL}(P_1, P_2)$, exists. Furthermore, since the ρ_k and α_k can be obtained by solving linear systems in $\mid Q_{1,2} \mid$ variables, the proof also gives a polynomial algorithm for computing $d^{ps}_{KL}(P_1, P_2)$. However, the worst-case complexity is still rather high: If $\mid Q_i \mid \sim N$, then $Q_{1,2}$ can have N^2 states, and the computation of the ρ_k and α_k involve inversions of matrices of dimensions $N^2 \times N^2$. Thus, depending on what algorithms are used for matrix inversion, the overall complexity is between $O(N^5)$ and $O(N^6)$. Alternatively, one can also approximate $d^{ps}_{KL}(P_1, P_2)$ by computing $1/(n-1)E^{(n-1)}[\sum_{k=1}^{n-1} f((q_{1,k}, q_{2,k}, \sigma_k))]$ for a sufficiently large n. This can be done in time $O(n \mid Q_{1,2} \mid)$.

Proof (Proof of Proposition 2). We prove the results in the following table, where the numbers in parenthesis refer to sections of the proof:

	KL	L_1	L_2	L_∞
d^λ	y (1)	y (1)	y (1)	y (1)
d^∞	y (2)	y (2)	n (3)	n (3)
d^{ps}	n (5)	n (4)	n (4)	n (4)

In all cases the direction $d(P_1, P_2) = 0 \Leftarrow q_1 \overset{T}{\sim} q_2$ is immediate, since the distances are functions only of the distributions P_{q_i}. Thus, we have to check whether, conversely, $P_1 \neq P_2$ implies $d(P_1, P_2) > 0$.

(1): If $P_1 \neq P_2$, then there exists an n and $w \in \Sigma^n$ with $P_1(w) \neq P_2(w)$. Then for $X \in \{KL, L_1, L_2, L_\infty\}$ $d^{(n)}_X(P_1, P_2) > 0$, and hence $d^\lambda_X(P_1, P_2) > 0$.

(2): For $X \in \{KL, L_1\}$ one has for all n that $d_X^{(n+1)} \geq d_X^{(n)}$. For KL this Lemmalem:dKLngrowth (i), and for L_1 it can be easily seen directly from the definition. From this, similar as in (1), one obtains that for some n: $0 < d_X^{(n)}(q_1, q_2) \leq d_X^{\infty}(q_1, q_2)$.

(3): Figure 1 shows an LMC model $\mathcal{M}_\epsilon$ for a a sequence of tosses with a biased or unbiased ($\epsilon = 0$) coin. Let P_ϵ be the distribution defined by $\mathcal{M}_\epsilon$.

For $n \geq 1$ and $i \leq n$ let $\Sigma_i^{(n)}$ consist of the $\{H, T\}$-words with exactly i occurrences of H. Also, let $a = 0.5 + \epsilon$, and $b = 0.5 - \epsilon$. Then

$$\begin{aligned}(d_{L_2}^{(n)}(P_0, P_\epsilon))^2 &= \textstyle\sum_{i=0}^{n} \sum_{w \in \Sigma_i^{(n)}} (P_0(w) - P_\epsilon(w))^2 \\ &= \textstyle\sum_{i=0}^{n} \binom{n}{i} (0.5^n - a^i b^{n-i})^2 \\ &= \textstyle\sum_{i=0}^{n} \binom{n}{i} (0.5^{2n} - 0.5^{n-1} a^i b^{n-i} + a^{2i} b^{2(n-i)}) \\ &= 0.5^{2n} \cdot 2^n - 0.5^{n-1}(a+b)^n + (a^2 + b^2)^n \\ &= (0.5 + 2\epsilon^2)^n - 0.5^n\end{aligned}$$

For any $\epsilon < 0.5$, thus, $d_{L_2}^{(n)}(P_0, P_\epsilon) \to 0$ as $n \to \infty$, and so $d_{L_2}^{\infty}(P_0, P_\epsilon) = 0$.

For L_∞ one has $d_{L_\infty}^{(n)}(P_0, P_\epsilon) = a^n - 0.5^n$, which also goes to 0 as $n \to \infty$.

(4) This immediately follows from the fact that $d^{ps} \equiv 0$ whenever the $d^{(n)}$ are bounded, and that for every $p \geq 1$ and $n \geq 1$ $d_{L_p}^{(n)}() \leq 2$.

(5) Figure 2 shows a deterministic LMC $\mathcal{M}_{k,p}$ parameterized by k (length of an initial sequence of a-labeled states), and p (the indicated transition probability). At present we only need to consider the case $k = 1$ (we will revisit the models $\mathcal{M}_{k,p}$ again later). For the automata $\mathcal{M}_{1,p}$ then $P_{q_{init}}(a^\omega) = p$ and $P_{q_{init}}(aba^\omega) = 1 - p$. Let q_1, q_2 be the initial states in $\mathcal{M}_{1,p}$ and $\mathcal{M}_{1,p'}$, respectively. Then for all $n \geq 2$ $d_{KL}^{(n)}(q_1, q_2)$ is constant equal to $p \log(p/p') + (1-p)\log((1-p)/(1-p'))$, and hence $d_{KL}^{ps}(q_1, q_2) = 0$.

Proof (Proof of Proposition 3). We prove the results listed in the following tables, where the numbers in parenthesis refer to sections of the proof:

	KL	L_1	L_2	L_∞
d^λ	y (2)	y (1)	y (1)	y (1)
d^∞	n (3)	n (3)		
d^{ps}	y (4)			

	$d_{b,\lambda}$
$\lambda = 1$	n (5)
$\lambda < 1$	y (6)

In the present context n is used to index the sequence π_n, and we use k as the index for the finite dimensional distance functions $d^{(k)}$.

(1) For $X \in \{L_1, L_2, L_\infty\}$ one has the global bound $d_X^{(k)}(P_1, P_2) \leq 2$ for all k, P_1, P_2. Therefore, for any given $\delta > 0$ there exists a $k > 0$, such that $\sum_{j>k} \lambda^j d_X^{(j)}(P, P_n) < \delta/2$, regardless of n. For any word $w \in \Sigma^*$: $P_n(w) \to P(w)$ as $n \to \infty$. For $X \in \{L_1, L_2, L_\infty\}$ it then immediately follows that for all k $d_X^{(k)}(P_n, P) \to 0$ ($n \to \infty$). For sufficiently large n, thus, also $\sum_{j=1}^{k} \lambda^j d_X^{(j)}(P, P_n) < \delta/2$.

(2) This is analogous to part (1), but now using Lemma 1 (iib) to bound for sufficiently large k the tail-sums $\sum_{j>k} \lambda^j d_{KL}^{(j)}(P, P_n) < \delta/2$ (uniformly for all n, and also for the symmetric case $d_{KL}^{(j)}(P_n, P)$).

(3) Consider again the models $\mathcal{M}_\epsilon$ as defined by Figure 1, and let P_ϵ be the distribution defined by $\mathcal{M}_\epsilon$. For $w \in \{H,T\}^*$ let $\%H(w) \in [0,1]$ denote the relative frequency of occurrence of H in w. For $n > 0$, and $\delta > 0$ define

$$O_{k,\delta} := \{w \in \{H,T\}^k \mid \%H(w) \in [1/2-\delta, 1/2+\delta]\}$$

Then by the weak law of large numbers

$$\lim_{k\to\infty} P_0(O_{k,\epsilon/2}) = 1, \qquad \lim_{k\to\infty} P_\epsilon(O_{k,\epsilon/2}) = 0.$$

For d_{KL} and d_{L_1} we can bound

$$d^{(k)}_{KL}(P_0,P_\epsilon) \geq P_0(O_{k,\epsilon/2}) \log \tfrac{P_0(O_{k,\epsilon/2})}{P_\epsilon(O_{k,\epsilon/2})} + P_0(\Sigma^k \setminus O_{k,\epsilon/2} \log \tfrac{P_0(\Sigma^k \setminus O_{n,\epsilon/2})}{P_\epsilon(\Sigma^k \setminus O_{n,\epsilon/2})})$$
$$d^{(n)}_{L_1}(P_0,P_\epsilon) \geq \mid P_0(O_{k,\epsilon/2}) - P_\epsilon(O_{k,\epsilon/2}) \mid + \mid P_0(\Sigma^k \setminus O_{k,\epsilon/2}) - P_\epsilon(\Sigma^k \setminus O_{k,\epsilon/2}) \mid$$

For d_{KL} this inequality is another consequence of the additivity property [9, Chapter 2], and for d_{L_1} it follows from basic arithmetic. As $k \to \infty$, the right hand sides of these inequalities go to ∞ (d_{KL}), respectively 2 (d_{L_1}). Thus, regardless of ϵ, $d^\infty_{KL}(P_0,P_\epsilon) = \infty$, and $d^\infty_{L_1}(P_0,P_\epsilon) = 2$.

(4) We show that $\lim_n \limsup_k 1/k\, d^{(k)}_{KL}(P,P_n) = \lim_n \limsup_k 1/k d^{(k)}_{KL}(P_n,P) = 0$. This shows parameter continuity of d^{ps}_{KL} whenever $d^{(k)}_{KL}(P,P_n)$ is well-defined for P and all P_n. Similar as in the proof of Lemma 1 we write:

$$d^{(k)}_{KL}(P,P_n) = \sum_{w\in\Sigma^k : P(w)>0} P(w) \log \frac{P(w)}{P_n(w)} = \sum_{w\in\Sigma^k : P(w)>0} P(w) \log \frac{\sum_{\boldsymbol{q}:L(\boldsymbol{q})=w} P(\boldsymbol{q})}{\sum_{\boldsymbol{q}:L(\boldsymbol{q})=w} P_n(\boldsymbol{q})}, \tag{10}$$

where $\boldsymbol{q} : L(\boldsymbol{q}) = w$ stands for all state sequences $\boldsymbol{q} \in Q^k$ whose label sequence is equal to w. Furthermore, we assume the summation to be restricted to those $\boldsymbol{q}$ which have a non-zero probability under P, or, equivalently, under P_n. Then we can bound the right-hand side of (10) by

$$\sum_{w\in\Sigma^k : P(w)>0} P(w) \log \max_{\boldsymbol{q}:L(\boldsymbol{q})=w} \frac{P(\boldsymbol{q})}{P_n(\boldsymbol{q})}. \tag{11}$$

For every $\delta > 0$ and all sufficiently large n we have $\frac{P(\boldsymbol{q})}{P_n(\boldsymbol{q})} \leq (1+\delta)^k$ for all $\boldsymbol{q}$ and k. Thus, (11) can be bounded by $k \log(1+\delta)$, and therefore $\limsup_k 1/k\, d^{(k)}_{KL}(P,P_n) \leq \log(1+\delta)$. The case for $d^{(k)}_{KL}(P_n,P)$ is analogous.

(5) As under (3), let q_1, q_2 be the H-states in $\mathcal{M}_0$, respectively $\mathcal{M}_\epsilon$. For any coupling $\mathcal{C}$ of $\mathcal{M}_0$ and $\mathcal{M}_\epsilon$, $\gamma^{\mathcal{C}}_1(q_1,q_2)$ is the probability that $\mathcal{C}$ starting from (q_1,q_2) will eventually reach a state labeled with (H,T) or (T,H) [3,1]. Since for any coupling the transition probability to (H,T) or (T,H) from both (H,H) and (T,T) is at least ϵ, this probability is one. Thus, $d_{b,1}(q_1,q_2) = 1$ for all $\epsilon > 0$.

(6) Let $\mathcal{M}, \mathcal{M}_n$ as in Definition 5. For clarity we denote in the following with q, q_n corresponding states of $\mathcal{M}$ and $\mathcal{M}_n$, i.e., q and q_n are the same elements

of Q, but equipped with the different transition probabilities π, respectively π_n. Given an $\epsilon > 0$, we can choose $n_0 \geq 1$ so that for all $n \geq n_0$, and all $q \in Q$: $\sum_{q' \in Q} |\pi(q,q') - \pi_n(q_n, q'_n)| < \epsilon$.

For $n \geq n_0$ we can now define couplings $\mathcal{C}_n$ of $\mathcal{M}$ and $\mathcal{M}_n$ so that for all q

$$\sum_{q' \in Q} \omega_n((q,q_n)(q',q'_n)) \geq 1 - \epsilon.$$

With $\gamma^{\mathcal{C}}_\lambda() \leq 1$, we then obtain for all $q \in Q$:

$$\begin{aligned}\gamma^{\mathcal{C}_n}_\lambda(q,q_n) &= \lambda(\sum_{q'} \gamma^{\mathcal{C}_n}_\lambda(q',q'_n)\omega_n((q,q_n)(q',q'_n)) + \sum_{s,t:s\neq t} \gamma^{\mathcal{C}}_\lambda(s,t'_n)\omega_n((q,q_n)(s,t'_n))) \\ &\leq \lambda(\max_{q'} \gamma^{\mathcal{C}_n}_\lambda(q',q'_n)(1-\epsilon) + \epsilon),\end{aligned}$$

and hence

$$\max_q \gamma^{\mathcal{C}_n}_\lambda(q,q_n) \leq \lambda(\max_q \gamma^{\mathcal{C}_n}_\lambda(q,q_n)(1-\epsilon) + \epsilon).$$

Since ϵ can be chosen arbitrarily small, this implies that $\max_q \gamma^{\mathcal{C}_n}_\lambda(q,q_n) \to 0$ for $n \to \infty$, and therefore also $d_{b,\lambda}(q,q_n) \to 0$ for all q.

Proof (Proof of Proposition 4). We prove the results in the following tables, where the numbers in parenthesis refer to sections of the proof.

	KL	L_1	L_2	L_∞
d^λ	y (1)	y (1)	y (1)	y (1)
d^∞	uy (2)	uy (2)		
d^{ps}	n (3)			

	$d_{b,\lambda}$
$\lambda = 1$	uy (4)
$\lambda < 1$	y (5)

(1) Let $X \in \{KL, L_1, L_2, L_\infty\}$, $w\Sigma^\omega$ be a cylinder set with $w \in \Sigma^n$, and $\epsilon > 0$. If $d^\lambda_X(P_1,P_2) < \delta$, then $d^{(n)}_X(P_1,P_2) < \delta/\lambda^n =: \gamma$. Furthermore, for δ and hence γ sufficiently small, $d^{(n)}_X(P_1,P_2) < \gamma$ implies $\mid P_1(w) - P_2(w) \mid < \epsilon$ (the exact dependence of γ on ϵ varies for $X = KL, L_1, L_2, L_\infty$).

That the d^λ_X are not uniformly BLTL continuous is demonstrated by the example of the automata $\mathcal{M}_{k,p}$ shown in Figure 2: the distance between the initial states of $\mathcal{M}_{k,p}$ and $\mathcal{M}_{k,0.5}$ decreases on the order of λ^k, but for all k $|P_{\mathcal{M}_{k,p},q_{init}}(\bigcirc^k b) - P_{\mathcal{M}_{k,p},q_{init}}(\bigcirc^k b)| = |p - 0.5|$.

(2) For $X \in \{KL, L_1\}$ one has that $d^{(n)}_X(P_1,P_2) \leq d^\infty_X(P_1,P_2)$ for all n. Thus, a bound on d^∞_X uniformly bounds all $d^{(n)}_X$, and hence, as above in part (1), all $\mid P_1(w) - P_2(w) \mid$.

(3) This follows by Lemma 4 and Proposition 2.

(4) This follows from Lemma 10 in [3].

(5) We first need to introduce some notation related to state traces in couplings $\mathcal{C}$. We denote with $(\boldsymbol{q}, \boldsymbol{q}') = (q_0, q'_0), ..., (q_{n-1}, q'_{n-1})$ finite traces in $\mathcal{C}$. We write $Tr((q,q'),n)$ for the set of all traces of length n starting at $(q_0, q'_0) = (q, q')$. The transition probabilities ω then also define a distribution over $Tr((q,q'),n)$. We denote with $LC((q,q'),n)$ the set of *label consistent* traces in $Tr((q,q'),n)$,

i.e., those $(\boldsymbol{q}, \boldsymbol{q}')$ with $L(q_i) = L(q'_i)$ for $i = 0, \ldots, n-1$. Finally, for $n \geq 1$ we define the event

$$(q, q') \stackrel{n}{\to} \textit{l.i.} := LC((q, q'), n-1) \setminus LC((q, q'), n).$$

Thus, $(q, q') \stackrel{n}{\to} \textit{l.i.}$ is the set of traces starting with (q, q') that become label inconsistent after exactly $n-1$ steps. Note that $(q, q') \stackrel{n}{\to} \textit{l.i.}$ can be seen as a subset of any $Tr((q, q'), m)$ with $m \geq n$, and that the value of $\omega((q, q') \stackrel{n}{\to} \textit{l.i.})$ does not depend on which $Tr((q, q'), m)$ is assumed as the underlying probability space.

The following claim is a generalization for the discounted case of a claim (somewhat implicitly, and without proof) already made in [3].

Claim 1:

Let $\mathcal{C} = \langle Q \times Q, \Sigma \times \Sigma, \omega, L \rangle$ be the coupling for which $d_{b,\lambda} = \gamma^{\mathcal{C}}_{\lambda}$. Then for all $n \geq 1$:

$$d_{b,\lambda}(q, q') \geq \sum_{i=0}^{n} \lambda^i \omega((q, q') \stackrel{i+1}{\to} \textit{l.i.}) \tag{12}$$

Proof of Claim 1: First, we note that the claim is true if $L(q) \neq L(q')$: then $d_{b,\lambda}(q, q') = 1$, $\omega((q, q') \stackrel{1}{\to} \textit{l.i.}) = 1$, and $\omega((q, q') \stackrel{i}{\to} \textit{l.i.}) = 0$ for all $i > 1$.

The case $L(q) = L(q')$ is by induction on n. For $n = 0$ we then have $\omega((q, q') \stackrel{1}{\to} \textit{l.i.}) = 0$, and so the right side of (12) evaluates to 0. For $n > 0$ we obtain:

$$\begin{aligned}
d_{b,\lambda}(q, q') &= \lambda \sum_{u,v} d_{b,\lambda}(u, v)\omega((q, q'), (u, v)) \\
&= \lambda \left(\omega((q, q') \stackrel{2}{\to} \textit{l.i.}) + \sum_{u,v:L(u)=L(v)} d_{b,\lambda}(u, v)\omega((q, q'), (u, v) \right) \\
&\geq \lambda \left(\omega((q, q') \stackrel{2}{\to} \textit{l.i.}) + \sum_{u,v:L(u)=L(v)} \sum_{j=1}^{n-1} \lambda^j \omega((u, v) \stackrel{j+1}{\to} \textit{l.i.})\omega((q, q'), (u, v)) \right) \quad (13) \\
&= \lambda \left(\omega((q, q') \stackrel{2}{\to} \textit{l.i.}) + \sum_{j=1}^{n-1} \lambda^j \omega((q, q') \stackrel{j+2}{\to} \textit{l.i.}) \right) \\
&= \sum_{i=1}^{n} \lambda^i \omega((q, q') \stackrel{i+1}{\to} \textit{l.i.}) = \sum_{i=0}^{n} \lambda^i \omega((q, q') \stackrel{i+1}{\to} \textit{l.i.})
\end{aligned}$$

For (13) the induction hypothesis is used, and the fact that because of $L(u) = L(v)$ we have $\omega((u, v) \stackrel{1}{\to} \textit{l.i.}) = 0$.

From claim 1 we obtain the slightly weaker statement

$$d_{b,\lambda}(q, q') \geq \lambda^n (1 - \omega(LC((q, q'), n+1))). \tag{14}$$

Now consider $w \in \Sigma^n$. By condition *1.* in the definition of couplings, we can write

$$P_q(w) = \sum_{\substack{(\boldsymbol{q},\boldsymbol{q}') \in Tr((q,q'),n):\\ L(\boldsymbol{q})=w}} \omega((\boldsymbol{q},\boldsymbol{q}')), \qquad P_{q'}(w) = \sum_{\substack{(\boldsymbol{q},\boldsymbol{q}') \in Tr((q,q'),n):\\ L(\boldsymbol{q}')=w}} \omega((\boldsymbol{q},\boldsymbol{q}')).$$

Thus, using (14):

$$|\, P_q(w) - P_{q'}(w) \,| \leq 1 - \omega(LC((q,q'),n)) \leq d_{b,\lambda}(q,q')/\lambda^{n-1}$$

Proof (Proposition 5). We prove the results in the following tables, where the numbers in parenthesis refer to sections of the proof:

	KL	L_1	L_2	L_∞
d^λ	n (1)	n (1)	n (1)	n (1)
d^∞	uy (2)	uy (2)		
d^{ps}	n (4)			

	$d_{b,\lambda}$
$\lambda = 1$	uy (3)
$\lambda < 1$	n (1)

(1) Consider again the automata $\mathcal{M}_{k,p}$ of Figure 2, and the LTL sentence $\Diamond b$ (eventually b). Let $p_1 \neq p_2$, and let $P_{i,k}$ be the distribution defined by the initial state of $\mathcal{M}_{k,p_i}$ $(i = 1, 2)$ Then, for all k: $|\, P_{1,k}(\Diamond b) - P_{2,k}(\Diamond b) \,| = |\, p_1 - p_2 \,|$. On the other hand, for all discounted distances $\lim_{k\to\infty} d(P_{1,k}, P_{2,k}) = 0$.

(2) Recall from Section 2 that $\mathcal{O}$ and $\mathcal{A}$ denote the class of open, respectively measurable, subsets of Σ^ω. We can actually show the much stronger statement that d^∞_{KL} and $d^\infty_{L_1}$ are uniformly $\mathcal{A}$-continuous. For this, we first show the following strengthening of Proposition 4:

Claim 2: d^∞_{KL} and $d^\infty_{L_1}$ are uniformly $\mathcal{O}$-continuous.

Proof of Claim 2: Let $O \in \mathcal{O}$. O is the union $\cup_{i=0}^\infty C_i$ of cylinder sets C_i. Let $\epsilon > 0$, and consider two probability measures P_1, P_2. Let $0 < m < \infty$ be such that for $i = 1, 2$:

$$P_i(O) - P_i(\cup_{j=0}^m C_j) \leq \epsilon/3.$$

Let n be the maximal prefix-length defining one of the C_j $(0 \leq j \leq m)$, i.e.:

$$n := \max\{k \mid C_j = w\Sigma^\omega, w \in \Sigma^k, 0 \leq j \leq m\}.$$

Now define

$$O^n := \{w \in \Sigma^n \mid w\Sigma^\omega \subseteq \cup_{j=0}^m C_j\}$$

Then $P_i(O^n) = P_i(\cup_{j=0}^m C_j)$ (where, by a slight abuse of notation, $P_i(O^n)$ stands for $P_i(\{w\Sigma^\omega \mid w \in O^n\})$), and

$$|\, P_1(O) - P_2(O) \,| \leq |\, P_1(O) - P_1(O^n) \,| + |\, P_1(O^n) - P_2(O^n) \,| + |\, P_2(O) - P_2(O^n) \,|\,.$$

With $|\, P_1(O^n) - P_2(O^n) \,| \leq d^{(n)}_{L_1}(P_1, P_2) \leq d^\infty_{L_1}(P_1, P_2)$ we thus obtain $|\, P_1(O) - P_2(O) \,| \leq \epsilon$ if $d^\infty_{L_1}(P_1, P_2) < \epsilon/3$. This proves the claim for $d^\infty_{L_1}$. For d^∞_{KL} it then follows from Lemma 2. From Claim 2 we obtain the stronger result:

Claim 3: d^∞_{KL} and $d^\infty_{L_1}$ are uniformly $\mathcal{A}$-continuous.

To prove Claim 3, we just need the fact that a probability measure P on $\mathcal{A}(\Sigma^\omega)$ is *regular*, which means that for all measurable $A \in \mathcal{A}$:

$$P(A) = \inf\{P(O) \mid A \subseteq O \in \mathcal{O}\}$$

(see [4, Chapter 8.1]). With this, Claim 3 immediately follows from Claim 2.

(3) For $d_{b,1}$ uniform $\mathcal{A}$-continuity is given by Corollary 11 of [3].

(4) Follows from Proposition 4.

Proof (Proposition 6). We construct an open set O, such that for $\mathcal{M}_0$ and $\mathcal{M}_\epsilon$ from Figure 1: $P_0(O) < 1/4$, and $P_\epsilon(O) > 1/2$, for all $\epsilon > 0$.

For $w \in \Sigma^*$ let $\%H(w)$ denote the relative frequency of occurrences of T in w. For $i, n \in \mathbb{N}$ define

$$O_{i,n} := \{w\Sigma^\omega \mid w \in \Sigma^n : \%H(w) \geq 1/2 + 2^{-i}\}$$

By the law of large numbers, there exists for every i an $n(i) \in \mathbb{N}$, so that $P_\epsilon(O_{i,n(i)}) \geq 1/2$ for all $\epsilon > 2^{-i+1}$, and $P_0(O_{i,n(i)}) < 2^{-(i+2)}$. Now $O = \cup_i O_{i,n(i)}$ has the desired properties.

Proof (Proposition 7). Each $O_{i,n(i)}$ as defined in the proof of Proposition 6 is BLTL-definable. Thus, for every $\epsilon > 0$ there exists a BLTL-definable property ϕ with $|P_0(\phi) - P_\epsilon(\phi)| \geq 1/4$.

Proof (Lemma 5). All claims are immediate from the definitions. The preservation of parameter continuity is the only case for which the constraint $\sum_i \alpha_i = 1$ is required (although $\sum_i \alpha_i < \infty$ would also be sufficient).

Proof (Proposition 8). A single d_ϕ obviously is $\{\phi_i\}$-continuous. LTL continuity then follows with Lemma 5.

That each d_ϕ is parameter continuous can be seen from the automata-theoretic approach to verification [21]: $P_{\mathcal{M},q}(\phi)$ can be computed as a reachability probability in the automaton constructed as the product of the automaton $\mathcal{M}$ containing q, and the automaton recognizing ϕ. The transition probabilities, and hence reachability probabilities, in the product are a continuous function of the transition probabilities in $\mathcal{M}$ (cf. also Theorem 2 of [11]).

References

1. G. Bacci, G. Bacci, K. G. Larsen, and R. Mardare. On-the-fly exact computation of bisimilarity distances. In *Proc. of TACAS 2013*, volume 7795 of *LNCS*, pages 1–15, 2013.
2. Christel Baier, EdmundM. Clarke, Vasiliki Hartonas-Garmhausen, Marta Kwiatkowska, and Mark Ryan. Symbolic model checking for probabilistic processes. In Pierpaolo Degano, Roberto Gorrieri, and Alberto Marchetti-Spaccamela, editors, *Automata, Languages and Programming*, volume 1256 of *Lecture Notes in Computer Science*, pages 430–440. Springer Berlin Heidelberg, 1997.

3. D. Chen, F. van Breugel, and J. Worrell. On the complexity of computing probabilistic bisimilarity. In *Proceedings of International Conference on Foundations of Software Science and Computation Structures (FoSSaCS)*, pages 437–451, 2012.
4. D. Cohn. *Measure Theory.* Birkhäuser, 1993.
5. Corinna Cortes, Mehryar Mohri, Ashish Rastogi, and Michael Riley. On the computation of the relative entropy of probabilistic automata. *Int. J. Found. Comput. Sci.*, 19(1):219–242, 2008.
6. Josee Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Metrics for labeled markov systems. In *CONCUR*, pages 258–273, 1999.
7. M. N. Do. Fast approximation of kullback-leibler distance for dependence trees and hidden markov models. *IEEE Signal Processing Letters*, 10(4):115–118, 2003.
8. R. M. Gray. *Entropy and Information Theory.* Springer, second edition edition, 2011.
9. S. Kullback. *Information Theory and Statistics.* Wiley, 1959.
10. Kim Guldstrand Larsen and Arne Skou. Bisimulation through probabilistic testing. *Inf. Comput.*, 94(1):1–28, 1991.
11. H. Mao, Y. Chen, M. Jaeger, T. D. Nielsen, K. G. Larsen, and B. Nielsen. Learning probabilistic automata for model checking. In *Proceedings of the 8th International Conference on Quantitative Evaluation of SysTems (QEST)*, pages 111–120, 2011.
12. J. R. Norris. *Markov Chains.* Cambridge University Press, 1997.
13. Z. Rached, F. Alajaji, and L. L. Campbell. The kullback-leibler divergence rate between markov sources. *IEEE Transactions on Information Theory*, 50(5):917–921, 2004.
14. K. Sen, M. Viswanathan, and G. Agha. Learning continuous time markov chains from sample executions. In *Proceedings of International Conference on Quantitative Evaluatio n of Systems (QEST)*, pages 146–155, 2004.
15. P. C. Shields. Two divergence-rate counterexamples. *Journal of Theoretical Probability*, 6(3):521–545, 1993.
16. J. Silva and S. Narayanan. Upper bound kullback-leibler divergence for transient hidden markov models. *IEEE Transactions on Signal Processing*, 56(9):4176–4188, 2008.
17. W. Thomas. Automata on infinite objects. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume 2. Elsevier/MIT Press, 1990.
18. G.T. Toussaint. Sharper lower bounds for discrimination information in terms of variation (corresp.). *Information Theory, IEEE Transactions on*, 21(1):99–100, 1975.
19. F. van Breugel, B. Sharma, and J. Worrell. Approximating a behavioural pseudometric without discount for probabilistic systems. *Logical Methods in Computer Science*, 4(2):1–23, 2008.
20. F. van Breugel and J. Worrell. A behavioural pseudometric for probabilistic transition systems. *Theoretical Computer Science*, 331:115–142, 2005.
21. M. Y. Vardi. Probabilistic linear-time model checking: an overview of the automata-theoretic approach. In J-P Katoen, editor, *Formal methods for real-time and probabilstic systems (ARTS-99) : 5th International AMAST Workshop*, volume 1601 of *LNCS*, 1999.