On approximation metrics for linear temporal model-checking of stochastic systems*

Ilya Tkachev
Delft University of Technology
i.tkachev@tudelft.nl

Alessandro Abate
University of Oxford
aabate@cs.ox.ac.uk

ABSTRACT

This paper proposes criteria for metrics between stochastic systems with a focus on the task of linear temporal model-checking. It explicitly puts forward two metrics which partially satisfy those criteria, and discusses their connection with other metrics studied in the literature. In particular, the notion of coupling between stochastic processes is shown to be crucial: omitting the explicit choice of coupling may lead to conservative results. The theoretical claims in the paper are supported by numerical examples.

1. INTRODUCTION

Stochastic systems have found broad applications in diverse areas where uncertainty can be quantified (cf. references in [31]). One particularly interesting class of problems concerns linear temporal (LT) model-checking of stochastic systems [8], which seeks to find the expectation of a path-dependent reward (or cost) functional, e.g. the probability that a realization of a system satisfies a given specification. Among the specifications of interest are reachability, safety, reach-avoid, and richer properties over a trajectory. If the system allows for a control input, one may further be interested in optimizing such an expectation or probability over all the admissible control policies [32]. Clearly, the simpler the system the easier the resulting model-checking procedure. In particular, if the state space of the process is finite, then model-checking can be performed algorithmically, by means of a specialized software [18, 21]. It is thus of interest to develop metrics[1] between stochastic systems that allow quantifying an error over LT properties when substituting a complex, concrete system with a simpler, abstract one.

A recent survey on stochastic metrics can be found in [3]. Here we briefly recapitulate two main approaches to defining and using such metrics. The first one has been taken in the Computer Science community, whereas the second has been proposed in the Systems & Control area.

The work in [11] has been the first to emphasize the necessity for metrics between stochastic systems, justifying their relevance by the conservative nature and lack of robustness of the notion of exact probabilistic bisimulation relation [22]. According to the notion in [22], an infinite system is bisimilar to a finite one only if the former has a certain "lumpable" structure relating the two models, and any small perturbation of the models can violate such a structure, thus resulting in systems that originally were precisely related and that become now completely unrelated. In contrast, the metric introduced in [11] both admits the bisimulation as its zero level set, and varies continuously with respect to perturbations of the system dynamics. A few other metrics have been developed along the same lines (see e.g. [35] and references therein), however in all these cases the applicability to LT model-checking has not been discussed.

Similar ideas have been later applied to non-stochastic systems [16]. The approach there is different from the one in [11, 35], as it starts from trace-like metrics over the global dynamics (paths), and then relates the former metrics on path spaces to a bisimulation-like metric over the local dynamics (transitions), the latter being more conservative but easier to compute. The developed metrics were designed to quantify errors over LT properties. The extension of this technique to stochastic systems has been proposed in [19], however the obtained results appear to be rather conservative, possibly due to the fact that the work with stochastic systems has been done at the level of random elements (rather than distributions), and as their coupling procedure is not addressed explicitly (cfr. Sections 2.1 and 3.2).

In summary, there are two main approaches to metrics between stochastic systems that are currently available. The first, taken in the Computer Science community [35, 11], does not provide an explicit way of using metrics for LT model-checking. The second, originated in the area of Systems & Control [19], although applicable to the aforementioned task, leads to conservative results. These reasons suggest to look into alternative approaches to define metrics between stochastic systems. We propose the following criteria for such metrics:

i. a metric between two systems quantifies the difference between LT specifications computed over such systems;

ii. given any two systems the metric between them is com-

*A. Abate is also affiliated with the Delft University of Technology. This work is supported by the European Commission STREP project MoVeS 257005, by the European Commission Marie Curie grant MANTRAS 249295, by the European Commission IAPP project AMBI 324432, by the European Commission NoE Hycon2 257462, and by the NWO VENI grant 016.103.020.

[1] In this paper we mostly deal with pseudo-metrics, which are allowed to be equal to zero for two different elements that are equivalent, in a certain sense. Since the difference between a metric and a pseudo-metric is not of importance to us, we refer to both objects simply as metrics.

HSCC'14, April 14–17, 2014, Berlin, Germany.

putable, either analytically or at least by means of Monte Carlo simulations;

iii. there is a class of "nice" infinite systems that allow for finite abstractions that are arbitrarily close in a given metric.

With focus on the criteria above, in this work we provide two new metrics. Unlike [19], we do not assume any form of stability over the systems under study, hence most of the results in this work hold true over a finite time horizon n, the dependence on which we then explicitly mention in the notation for the metrics. The first metric d_{TV^n} is based on the total variation distance between measures, whereas the second metric d_{W^n} is inspired by the Wasserstein distance.[2] Similar to [16, 19], we start with trace-like metrics to assure that the first criterion is satisfied. We also provide Monte Carlo methods to quantify both metrics in order to meet the requirements of the second criterion. Finally, we provide precise (rather than simulation-based) bounds on d_{TV^n} in terms of a bisimulation-like analogue, here called d_{B}: this allows us to characterize a class of systems that satisfy the last criterion with respect to d_{TV^n}, which in turn aligns with results on formula-free abstractions proposed in [31]. Unfortunately, such analytic results are much harder to obtain for the metric d_{W^n}.

The rest of the work is structured as follows. The majority of our ideas and methods rely upon the notion of random elements, their distributions, and their couplings. We briefly go over this theory in Section 2, which also introduces labelled discrete-time Markov processes (ldt-MP), the class of systems we focus on in this paper. The metrics d_{TV^n} and d_{W^n}, and relations between them, are discussed in Section 3. Theoretical and computational examples are further provided in Section 4, and the paper is concluded in Section 5. The notation can be found in Section 7.

2. PRELIMINARIES

2.1 Coupling

The notion of coupling [23] is crucial when considering an interplay between two stochastic processes, particularly in case one wants to quantify the difference between them. To introduce this notion, we need to elaborate on the definition of stochastic processes and on their representations. Recall that a stochastic process is a special case of a random element – this concept is often used to model probabilistic phenomena.

DEFINITION 1. *A* random element *on a measurable space $(E,\mathscr{E})$ is a tuple $(\Omega,\mathscr{F},\mathsf{P},f)$, where $(\Omega,\mathscr{F},\mathsf{P})$ is a probability space and $f:\Omega\to E$ is a measurable map. We say that $(\Omega,\mathscr{F},\mathsf{P})$ is the* sample space *of the random element, and that $(E,\mathscr{E})$ is its* range. *The* distribution *of the random element $(\Omega,\mathscr{F},\mathsf{P},f)$ is the probability measure $f_*\mathsf{P}\in\mathscr{P}(E,\mathscr{E})$ on its range.*

We will shortly discuss that distributions can be induced by different random elements, thus one can think of the latter as particular *representations* of distributions. More precisely, we say that a random element $(\Omega,\mathscr{F},\mathsf{P},f)$ is the representation of a probability measure Q whenever it holds that $f_*\mathsf{P}=\mathsf{Q}$. Note that, for any distribution $\mathsf{Q}\in\mathscr{P}(E,\mathscr{E})$, there always exists at least one representation over its range, given by $(E,\mathscr{E},\mathsf{Q},\mathrm{id}_E)$ since $(\mathrm{id}_E)_*\mathsf{Q}=\mathsf{Q}$. This is also known as the *canonical representation* of the distribution Q[3]. Let us provide some examples to further clarify the concept of representation. The first example shows that for any distribution a representation is never unique.

EXAMPLE 1. *Consider a distribution Q and let $(\Omega,\mathscr{F},\mathsf{P},f)$ be its arbitrary representation. For any probability space $(\Omega',\mathscr{F}',\mathsf{P}')$ it holds that $(\Omega\times\Omega',\mathscr{F}\otimes\mathscr{F}',\mathsf{P}\otimes\mathsf{P}',f\circ\pi)$ is another representation of Q, where $\pi:\Omega\times\Omega'\to\Omega$ denotes the projection map.*

In the previous example it is crucial that the sample space is modified. The next example emphasizes this fact by showing that if the sample space is fixed, then there may exist distributions that admit a unique representation over this sample space.

EXAMPLE 2. *Let $E:=\{-1,1\}$, $\mathscr{E}:=2^E$, and let $\mathsf{Q}(\{1\})=\frac{1}{3}$. Note that there exist four distinct maps $f:E\to E$, that is id_E, $-\mathrm{id}_E$, and the constant maps -1 and 1. Clearly, $\mathsf{Q}=(\mathrm{id}_E)_*\mathsf{Q}$ but $\mathsf{Q}\neq f_*\mathsf{Q}$ if $f\neq\mathrm{id}_E$, so that Q can be represented in the unique (canonical) way if the sample space is $(E,\mathscr{E})$. On the other hand, a symmetrical distribution Q' given by $\mathsf{Q}'(\{1\})=\frac{1}{2}$ can be represented in two ways since $\mathsf{Q}'=(-\mathrm{id}_E)_*\mathsf{Q}'$. More generally, if a distribution admits distinct representations over its range, they can be considered as symmetries of this distribution.*

According to [24, Section 2.1], a stochastic process on a measurable space $(X,\mathscr{X})$ is a parameterized collection of X-valued random elements $(\mathbf{x}_t)_{t\in T}$, all defined over the same probability space $(\Omega,\mathscr{F},\mathsf{P})$. Clearly, one can equivalently consider a stochastic process as a single random element $(\Omega,\mathscr{F},\mathsf{P},f)$ where the map $f:\Omega\to X^T$ is uniquely determined by $f\circ\pi_t=\mathbf{x}_t$ for any t, $\pi_t:X^T\to X$ being obvious projection maps. Recall that here X^T is the product measurable space consisting of all maps from T to X. As a consequence, all the results on random elements above now apply to stochastic processes. In particular, it is important for us that any stochastic process can be understood also in a "weak" form – as a distribution over X^T, rather than only in a perhaps more intuitive "strong" form – as a random element inducing such a distribution. This approach is important in the context of this work, since only distributions matter for LT model checking, rather than particular representations of a stochastic process. A representation, especially if it is not canonical, is sometimes provided to show a constructive definition of a distribution, as in the following example.

EXAMPLE 3. *A one-dimensional Itô diffusion [24, Chapter 7]*

$$\mathrm{d}\mathbf{x}_t=a(\mathbf{x}_t)\mathrm{d}t+b(\mathbf{x}_t)\mathrm{d}B_t,\quad \mathbf{x}_0=x\in\mathbb{R}, \tag{1}$$

is a stochastic process with trajectories in $E=\mathscr{C}([0,\infty))$. Since the diffusion (1) is Markovian, it can be considered from two perspectives: either as a solution of the stochastic differential equation (SDE) (1), or as a Markov process on $X=\mathbb{R}$.

In order to treat a diffusion as a solution of an SDE, we consider a probability space $(\Omega,\mathscr{B}(\Omega),\mathsf{P})$, where $\Omega=\mathscr{C}_0([0,\infty))$ is a set of all $\omega\in E$ satisfying $\omega(0)=0$ and where P is the Wiener measure (the distribution of the Brownian motion $(B_t)_{t\in[0,\infty)}$). In this setting a diffusion is constructed as a random element $(\Omega,\mathscr{F},\mathsf{P},f)$, where f is the strong solution *of an SDE [24, Section 5.3] with the initial condition x. Let us denote the induced distribution by $\mathsf{Q}:=f_*\mathsf{P}$.*

[2] The Wasserstein metric is also known as Kantorovich or Hutchinson metric [35],[14, footnote in Section 2.2].

[3] The existence of a canonical representation in particular means that in order to represent any given distribution its range suffices and there is no necessity in coming up with a "bigger" sample space $(\Omega,\mathscr{F})$.

Alternatively, one can treat the diffusion in (1) as a Markov process, and construct a distribution Q directly on E using the transition function associated with the diffusion [13, Section 4.1]. The stochastic process for the diffusion (in a strong sense) would be a canonical representation of Q. In this Markovian setting there is no need to define an auxiliary probability space $(\Omega, \mathscr{F}, \mathsf{P})$ for the Brownian motion, so the current Markovian approach is more direct. On the other hand, such a construction is perhaps less intuitive as it does not emphasize the connection between the diffusion and the Brownian motion it is driven by.

Note that both methods lead to the same stochastic process in a weak sense (at the level of distributions), but to different stochastic processes in a strong sense (at the level of random elements). At the same time, if one needs to compute the probability that a trajectory of (1) reaches a target set, it does not matter which representation of a diffusion is used, since such a probability is uniquely determined by the distribution of the diffusion.

The motivation behind approximate abstractions of stochastic systems is to use the information obtained over a simpler abstract system to deduce properties of a more complicated concrete system. Assume that the range $(E, \mathscr{E})$ is endowed with a metric d_E. Given a concrete random element $(\Omega, \mathscr{F}, \mathsf{P}, f)$ and its approximation $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathsf{P}}, \tilde{f})$ one wants to obtain an inequality

$$\mathbb{P}\left(d_E(f, \tilde{f}) > \delta\right) \leq \varepsilon \tag{2}$$

to be able to interpret over f the results obtained for $\tilde{f}$. As an example [6], if we know the probability $p = \tilde{\mathsf{P}}(\tilde{f} \in A)$ for some set $A \in \tilde{\mathscr{F}}$, we can use (2) to provide bounds on $\mathsf{P}(f \in A_\delta)$, where $A_\delta = \{x \in E : d_E(x, A) \leq \delta\}$ is the δ-inflation of the set A:

$$\begin{aligned}\mathsf{P}(f \in A_\delta) = \mathbb{P}(f \in A_\delta) &\geq \mathbb{P}\left(\{\tilde{f} \in A\} \cap \{d_E(f, \tilde{f}) \leq \delta\}\right) \\ &\geq \mathbb{P}(\tilde{f} \in A) - \mathbb{P}(d_E(f, \tilde{f}) > \delta) \geq p - \varepsilon.\end{aligned}$$

Whenever f and $\tilde{f}$ are stochastic processes and A is a reachability specification, the result allows using the probabilistic reachability analysis over $\tilde{f}$ to study that over f. Recall that the abstraction is a random element on its own, and that the analysis over the abstraction can be carried out regardless of its relation to the concrete system: such a relation only matters when extrapolating results of this analysis from the abstraction back to the concrete system. For this purpose one has to define a *common* sample space for both systems in order to specify the probability measure in (2). The procedure of building a common sample space for *a priori* unrelated distributions or random elements is called coupling.

Definition 2. *A* coupling *of $\mathsf{Q}, \tilde{\mathsf{Q}} \in \mathscr{P}(E, \mathscr{E})$ is a probability measure $\mathbb{Q} \in \mathscr{P}(E^2, \mathscr{E}^2)$ satisfying the following equalities:*

$$\pi_* \mathbb{Q} = \mathsf{Q}, \qquad \tilde{\pi}_* \mathbb{Q} = \tilde{\mathsf{Q}}, \tag{3}$$

where $\pi(x, \tilde{x}) = x$ and $\tilde{\pi}(x, \tilde{x}) = \tilde{x}$ for all $(x, \tilde{x}) \in E^2$ are obvious projection maps. A coupling of two random elements $(\Omega, \mathscr{F}, \mathsf{P}, f)$ and $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathsf{P}}, \tilde{f})$ is any random element $(\Xi, \mathscr{G}, \mathbb{P}, (g, \tilde{g}))$ such that $(g, \tilde{g})_* \mathbb{P}$ is a coupling of $f_* \mathsf{P}$ and $\tilde{f}_* \tilde{\mathsf{P}}$, and $(\Xi, \mathscr{G})$ is an arbitrary measurable space.

Whenever (3) holds true, we say that Q and $\tilde{\mathsf{Q}}$ are *marginals* of $\mathbb{Q}$. It shall be clear that given any two marginal distributions, there always exists at least one coupling of them, called the *independence coupling*, which is given by $\mathbb{Q} := \mathsf{Q} \otimes \tilde{\mathsf{Q}}$. The most important point about the coupling is that it is only unique when one of the marginals Q or $\tilde{\mathsf{Q}}$ is the Dirac measure, or equivalently when one of the random elements is deterministic. In particular, since the concrete system and the abstraction are almost never coupled a priori (neither at the level of random elements, nor at the level of distributions), one can e.g. optimize over all admissible couplings to choose the best for inequality (2). This idea constitutes to the core of our method. Before going into the details of it, let us provide examples of couplings.

Example 4. *Consider the diffusion $\mathbf{x}$ as per (1), and let*

$$\mathrm{d}\tilde{\mathbf{x}}_t = \tilde{a}(\tilde{\mathbf{x}}_t)\mathrm{d}t + \tilde{b}(\tilde{\mathbf{x}}_t)\mathrm{d}\tilde{B}_t, \quad \tilde{\mathbf{x}}_0 = \tilde{x} \in \mathbb{R}, \tag{4}$$

be another diffusion. Let us represent both of them as solutions of SDEs, that is as random elements on the sample spaces $(\Omega, \mathscr{F}, \mathsf{P})$ and $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathsf{P}})$, where the former is defined in Example 3, and the latter is defined for (4) analogously. We provide three versions of couplings obtained via the Brownian motions driving these diffusions:

a. $\Xi = \Omega$, $\mathscr{G} = \mathscr{F}$, $\mathbb{P} = \mathsf{P}$ and $(g, \tilde{g}) = (f, \tilde{f})$. This coupling corresponds to the case $B_t = \tilde{B}_t$ when diffusions are "sharing" the same noise [19].

b. $\Xi = \Omega \times \tilde{\Omega}$, $\mathscr{G} = \mathscr{F} \otimes \tilde{\mathscr{F}}$, $\mathbb{P} = \mathsf{P} \otimes \tilde{\mathsf{P}}$ and $(g, \tilde{g}) = (f \circ \pi, \tilde{f} \circ \tilde{\pi})$ where the projection maps $\pi : \Xi \to \Omega$ and $\tilde{\pi} : \Xi \to \tilde{\Omega}$ are as in Definition 2. This clearly corresponds to the case of the independence coupling, that is $B_t \perp \tilde{B}_t$ [2].

c. $\Xi = \Omega$, $\mathscr{G} = \mathscr{F}$, $\mathbb{P} = \mathsf{P}$ and $(g, \tilde{g}) = (f, \tilde{f} \circ n)$ where $n(\omega) = -\omega$ for any $\omega \in \Omega$. In this case $B_t = -\tilde{B}_t$: noises driving diffusions are "reflected". Such a construction is possible thanks to the fact that n is a symmetry of P (i.e. $n_* \mathsf{P} = \mathsf{P}$), or in other words $-\tilde{B}_t$ is a Brownian motion whenever $\tilde{B}_t$ is.

Finally, let us mention that as much as any distribution of a single random element admits a representation over its range, the distribution of any coupling of two random elements admits a representation over the product of the ranges: this follows directly from Definition 2. In particular, if $(\Omega, \mathscr{F}, \mathsf{P}, f)$ and $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathsf{P}}, \tilde{f})$ are any random elements with a range $(E, \mathscr{E})$ endowed with a metric d, and $(\Xi, \mathscr{G}, \mathbb{P}, (g, \tilde{g}))$ is one of their couplings with a distribution $\mathbb{Q} = (g, \tilde{g})_* \mathbb{P}$, then the probability in (2) can be expressed as

$$\mathbb{P}\left(d(g, \tilde{g}) > \delta\right) = \mathbb{Q}(E^2 \setminus \Delta_E^\delta),$$

where $\Delta_E^\delta := \{(x, \tilde{x}) : d(x, \tilde{x}) \leq \varepsilon\}$ is a diagonal ε-strip in E^2. The benefit of dealing with stochastic systems (and their couplings) at the level of distributions lies the easy introduction of the distance between the concrete and the abstract systems. To support this point, let us mention that in probability theory distances are classically introduced between distributions, rather than between random elements [15, 25].

2.2 Behaviors of ldt-MP

Although most of the concepts and methods we have introduced apply to arbitrary stochastic processes, a focus on discrete-time Markov processes (dt-MP) allows us to provide a more detailed analysis. A dt-MP is a pair (X, P), where X is a Borel space referred to as the state space, and $P : X \to \mathscr{P}(X)$ is a stochastic kernel on X. It follows from [26, Theorem 2.8] that given any initial state $x \in X$ there exists a unique probability measure P_x on the state-path space X^ω satisfying

$$\mathsf{P}_x\left(\prod_{k=0}^{\infty} \mathrm{d}x_k\right) = 1_{\{x\}}(x) \cdot \prod_{k=0}^{\infty} P(x_k, \mathrm{d}x_k). \tag{5}$$

We say that (X,P) is finite whenever X is a finite set.

REMARK 1. *Similar to Example 3, for a given dt-MP and a fixed initial state, (5) provides a weak stochastic process. The corresponding canonical strong stochastic process can be introduced by defining the coordinate maps $\mathbf{x}_k : X^\omega \to X$ as obvious projections. As an alternative, it follows from [20, Proposition 7.6] that any dt-MP can be expressed as a stochastic difference equation $\mathbf{x}_{k+1} = F(\mathbf{x}_k, w_k)$, where w_k is a sequence of iid uniform random variables on $[0,1]$. With focus on Section 2.1, this means that any dt-MP with a fixed initial state $x \in X$ admits the following non-canonical representation: $([0,1]^\omega, \mathscr{B}([0,1]^\omega), \lambda^\omega, f)$, where λ is the Lebesgue measure on $[0,1]$ and the map f is obtained by iterating F starting from x.*

For a non-stochastic system on a state space X, its internal behavior is any element of X^ω where each transition $\mathbf{x}_k \to \mathbf{x}_{k+1}$ is allowed in the system [29, Section 1.2]. For stochastic systems such a definition is not suitable since formally all transitions are allowed, so that any single element of X^ω is a possible internal behavior of a dt-MP, albeit possibly of zero probability. Due to this reason, it is more appropriate to talk about *behavioral distributions*, that is P_x shall be understood as a distribution of internal behaviors that indicates which behaviors are more likely appear as trajectories of a dt-MP.

Often one is not interested in each single behavior, but rather in collections thereof that can be characterised by means of observations (or labels). A labeled dt-MP (ldt-MP for short) is a tuple (X,P,Y,L) where (X,P) is a dt-MP, Y is a Borel space and $\mathsf{L} : X \to Y$ is a measurable map [31]. In an ldt-MP, to any internal behavior $(x_0, x_1, \dots)$ there corresponds an *external behavior*, or a *trace*, given by an output of the trace map:

$$\mathsf{L}_\omega(x_0, x_1, \dots) := (\mathsf{L}(x_0), \mathsf{L}(x_1), \dots).$$

For non-stochastic system a collection of its all admissible external behaviors is also called a *language* of a system [17]. Similar to the case of internal behaviors, in our setting it is more appropriate to talk about distributions of external behaviors as all of them are allowed in the ldt-MP. We refer to such distributions as *trace distributions* [27]: let us emphasize again that trace distributions are analogues of languages for non-stochastic systems. Since the map $\mathsf{L}_\omega : X^\omega \to Y^\omega$ is measurable [31, Theorem 1], for any ldt-MP its trace distribution can be expressed as $(\mathsf{L}_\omega)_* \mathsf{P}_x$. We can formulate the LT model-checking problem for ldt-MP as follows:

PROBLEM 1. *Given an ldt-MP (X,P,Y,L) and the observation-path dependent functional $f \in \mathscr{B}(Y^\omega)$, compute the expectation $((L_\omega)_* \mathsf{P}_x) f$ for any initial state $x \in X$.*

An important case of a functional f in Problem 1 is the indicator function $f = 1_A$ of some event of interest $A \in \mathscr{B}(Y^\omega)$: in such a case the expectation to be computed turns out to be the probability $((L_\omega)_* \mathsf{P}_x)(A) = \mathsf{P}_x(L_\omega(\mathbf{x}) \in A)$. A common example of an observation space is given by finite sets Y, also called *alphabets*. Over alphabets the event A can be for instance an ω-regular language[4] expressed as an automaton, or an LTL formula – for a detailed exposition see e.g. [8, Chapters 4, 5].

Dealing with stochastic systems at the level of observations allows one to compare systems possibly endowed with different state spaces. This feature is extremely important since a complex ldt-MP can be approximated by a simpler one over a smaller state space – for example, by a finite ldt-MP: for the latter LT model-checking allows for analytical solutions, and numerical procedures can be computationally efficient. In order to be able to quantitatively argue about the original ldt-MP using results obtained over its abstract approximation, it is useful to endow the trace space with some metrics [16]. The choice of the latter depends on how one wants to interpret over the original system the results obtained over the abstraction. Along the lines of the discussion in Section 2.1, for ldt-MP we define these metrics to measure the difference (or similarity) between trace distributions, regardless of the way the latter are represented. The trace equivalence for any two ldt-MP can be defined by requiring them to have the same trace distributions, however there may be several choices of metrics whose zero level sets coincide with such trace equivalence. The next section proposes two such metrics for the trace distributions over the ldt-MP with the same observation spaces based on the *total variation* distance TV and on the *Wasserstein* distance W.

3. METRICS FOR LDT-MP

3.1 Total variation distance

Perhaps the most direct way to define a distance between two probability measures that fits the purposes of Problem 1 is to maximize over all functionals (or events) the difference between the corresponding expectations. Interestingly, such seemingly naïve approach yields a useful metric called the total variation distance (see Section 7.2).

Let us consider two ldt-MP $\mathscr{D} = (X,P,Y,\mathsf{L})$ and $\tilde{\mathscr{D}} = (\tilde{X},\tilde{P},Y,\tilde{\mathsf{L}})$ over the same observation space Y. Given initial states $x \in X$ and $\tilde{x} \in \tilde{X}$, we denote the trace distributions of $\mathscr{D}$ and $\tilde{\mathscr{D}}$ by Q_x and $\tilde{\mathsf{Q}}_{\tilde{x}}$, respectively. Suppose for example that $\tilde{\mathscr{D}}$ is finite, Y is a finite alphabet and Φ is some LTL formula over Y. If we know the value of $\mathsf{TV}(\mathsf{Q}_x, \tilde{\mathsf{Q}}_{\tilde{x}})$, we can compute $\tilde{\mathsf{Q}}_{\tilde{x}}(\Phi)$ and use it to estimate $\mathsf{Q}_x(\Phi)$, since by definition of TV (12)

$$|\mathsf{Q}_x(\Phi) - \tilde{\mathsf{Q}}_{\tilde{x}}(\Phi)| \le \frac{1}{2}\mathsf{TV}(\mathsf{Q}_x, \tilde{\mathsf{Q}}_{\tilde{x}}).$$

From the definition of TV (for details see Section 7.2), it follows that similar bounds in terms of $\mathsf{TV}(\mathsf{Q}_x, \tilde{\mathsf{Q}}_{\tilde{x}})$ can be also obtained on the difference of expectations for more general cost functionals, rather than just indicator functions of LTL formulae. As a result, the TV satisfies the first criterion from Section 1 and can thus represent a good candidate for a metric. However, notice that $\mathsf{TV}(\mathsf{Q}_x, \tilde{\mathsf{Q}}_{\tilde{x}})$ quantifies the distance between probabilities on the infinite time horizon: this may be a too conservative requirement, as mentioned in [33, Section 3.1] and as the following example shows.

EXAMPLE 5. *Let us consider two simple ldt-MP with only two states: $\tilde{X} = X = \{0,1\}$, $Y = X$, $\mathsf{L} = \tilde{\mathsf{L}} = \mathrm{id}_X$, and suppose that transition matrices $P \neq \tilde{P}$ have only positive entries. It follows that these ldt-MP are ergodic; we denote their unique invariant distributions by μ and $\tilde{\mu}$, respectively. For any $h : X^2 \to \mathbb{R}$ define*

$$A_h := \left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} h(\mathbf{x}_k, \mathbf{x}_{k+1}) = \int_{X^2} h(x,y) P(x, \mathrm{d}y) \mu(\mathrm{d}y) \right\}.$$

It holds that $\mathsf{P}_x(A_h) = 1$ for any initial state x. A similar fact can be obtained for $\tilde{\mathsf{P}}_x$. Clearly, we can always find an h such that

$$\int_{X^2} h(x,y) P(x, \mathrm{d}y) \mu(\mathrm{d}y) \neq \int_{X^2} h(x,y) \tilde{P}(x, \mathrm{d}y) \tilde{\mu}(\mathrm{d}y).$$

[4] Indeed, ω-regular languages over a finite Y are always elements of $\mathscr{B}(Y^\omega)$, for the proof see e.g. [36, Proposition 2.3].

Since $\mathsf{Q}_x = \mathsf{P}_x$ *and* $\tilde{\mathsf{Q}}_x = \tilde{\mathsf{P}}_x$ *we get* $\mathsf{TV}(\mathsf{Q}_x, \tilde{\mathsf{Q}}_x) = 2$ *by putting* A_h *in* (12)*, no matter how small the difference between* P *and* $\tilde{P}$ *is.*

Due to the reasons discussed above, we focus our attention to finite-horizon behaviors of ldt-MP, characterized by restrictions of the trace distribution Q_x and $\tilde{\mathsf{Q}}_{\tilde{x}}$ to the set Y^{n+1}: we denote them by Q^n_x and $\tilde{\mathsf{Q}}^n_{\tilde{x}}$, respectively. Since given an initial state over the concrete ldt-MP one has the freedom of choosing that over the abstraction, let us define the TV-like distance between ldt-MPs with the same observation spaces by

$$d_{\mathsf{TV}^n}(\mathscr{D}, \tilde{\mathscr{D}}) := \sup_{x \in X} \inf_{\tilde{x} \in \tilde{X}} \mathsf{TV}(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_{\tilde{x}}). \tag{6}$$

Definition 3. *We say that two ldt-MP* $\mathscr{D}$ *and* $\tilde{\mathscr{D}}$ *are* ε-trace equivalent *in the* d_{TV^n} *metrics if* $d_{\mathsf{TV}^n}(\mathscr{D}, \tilde{\mathscr{D}}) \leq \varepsilon$.

Although the approximate trace equivalence of ldt-MP in d_{TV^n} for $n < \infty$ is not in general useful in infinite-horizon LT model-checking, it is sufficient to argue about finite-horizon properties, for example those expressed as BLTL formulae [31, Section 2.4]. Until the end of Section 3.1 we focus exclusively on the case when Y is a finite alphabet. A finite alphabet on a finite time horizon contains only a finite number of elements, which justifies the use of the following formula for the total variation [9]:

$$\mathsf{TV}(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_{\tilde{x}}) = \sum_{y \in Y^{n+1}} |\mathsf{Q}_x(\{y\}) - \tilde{\mathsf{Q}}_{\tilde{x}}(\{y\})|. \tag{7}$$

Though the expression above is still challenging to compute precisely even over finite-state ldt-MPs, it can be computed by means of Monte Carlo simulations, which shows that d_{TV^n} satisfies the second criterion mentioned in Section 1.

Let us fix the time horizon n, and let us sample independently N copies of the observations of $\mathscr{D}$ and $\tilde{\mathscr{D}}$ over the given time horizon, which we further denote by $(\mathbf{y}^i)_{i=1}^N$ and $(\tilde{\mathbf{y}}^i)_{i=1}^N$ respectively.[5] The index i refers to different runs: each of them can be obtained e.g. by sampling the state-path of $\mathscr{D}$, say $\mathbf{x}^i$, on the time horizon n, and then by mapping $\mathbf{x}^i$ into $\mathbf{y}^i$ by means of map L. Define

$$\mathsf{Q}^{n,N}_x(\cdot) := \frac{1}{N}\sum_{i=1}^N 1\{\mathbf{y}^i \in \cdot\}, \qquad \tilde{\mathsf{Q}}^{n,N}_{\tilde{x}}(\cdot) := \frac{1}{N}\sum_{i=1}^N 1\{\tilde{\mathbf{y}}^i \in \cdot\} \tag{8}$$

to be empirical distributions. It is fairly simple to compute the approximate distance $\hat{v}_N := \mathsf{TV}(\mathsf{Q}^{n,N}_x, \tilde{\mathsf{Q}}^{n,N}_{\tilde{x}})$ using (7), which leads to assess how good such an approximation of the original distance $v := \mathsf{TV}(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_{\tilde{x}})$ is. Let us denote by Π the joint probability distribution of the two iid sequences $(\mathbf{y}^i)_{i=1}^N$ and $(\tilde{\mathbf{y}}^i)_{i=1}^N$.

Theorem 1. *For any* $\delta > 0$ *it holds that*

$$\Pi\left(|v - \hat{v}_N| \leq 2\delta\right) \geq \left(1 - (N+1)^{|Y^{n+1}|} \mathrm{e}^{-N\delta^2}\right)^2, \tag{9}$$

provided that $(N+1)^{|Y^{n+1}|}\mathrm{e}^{-N\delta^2} \leq 1$.

Proof. The idea of the proof is to start with $\mathsf{TV}(\mathsf{Q}^n_x, \mathsf{Q}^{n,N}_x)$:

$$\Pi\left(\mathsf{TV}(\mathsf{Q}^n_x, \mathsf{Q}^{n,N}_x) \geq \delta\right) = \sum_{\nu \in L^\delta_N} \Pi\left(\mathsf{Q}^{n,N}_x = \nu\right) \leq \sum_{\nu \in L^\delta_N} \mathrm{e}^{-N \cdot \mathsf{KL}(\nu, \mathsf{Q}^n_x)},$$

where by KL we denote the Kullback-Leibler divergence between probability measures [15], L^δ_N is the set of all empirical measures ν satisfying $\mathsf{TV}(\mathsf{Q}^n_x, \nu) \geq \delta$, and the last inequality immediately follows from [9, Lemma 2.1.9]. Since for all $\nu \in L^\delta_N$ it holds that $\mathsf{KL}(\nu, \mathsf{Q}^n_x) \geq (\mathsf{TV}(\nu, \mathsf{Q}^n_x))^2$, and [9, Lemma 2.1.2(a)] implies that $|L^\delta_N| \leq (N+1)^{|Y^{n+1}|}$, we obtain:

$$\Pi\left(\mathsf{TV}(\mathsf{Q}^n_x, \mathsf{Q}^{n,N}_x) \geq \delta\right) \leq (N+1)^{|Y^{n+1}|}\mathrm{e}^{-N\delta^2}. \tag{10}$$

A similar estimate obviously can be derived for $\mathsf{TV}(\tilde{\mathsf{Q}}^n_{\tilde{x}}, \tilde{\mathsf{Q}}^{n,N}_{\tilde{x}})$. By triangular inequality we further get

$$|v - \hat{v}_N| \leq \mathsf{TV}(\mathsf{Q}^n_x, \mathsf{Q}^{n,N}_x) + \mathsf{TV}(\tilde{\mathsf{Q}}^n_{\tilde{x}}, \tilde{\mathsf{Q}}^{n,N}_{\tilde{x}}),$$

which together with (10) yields (9) as desired. □

Clearly, (9) implies that the estimator $\hat{v}_N$ converges in Π to the true distance v, and further provides an explicit bound on the confidence level. Note also that for fixed sizes of the alphabet and of the time horizon, the bound in (9) depends quadratically on the precision δ, and logarithmically on the confidence level. Unfortunately, such a dependence is only asymptotical, and for N not sufficiently large the polynomial term $(N+1)^{|Y^{n+1}|}$ dominates the exponential term $\mathrm{e}^{-N\delta^2}$. Moreover, such polynomial term grows very fast with respect to the size of the alphabet, and even faster with respect to the time horizon. For example, even if $|Y| = 2$, $n = 9$ and $\delta = 0.1$, one needs approximately an order of $N = 1.5 \times 10^6$ samples to obtain good confidence levels, whereas for $n = 10$ around $N = 3.5 \times 10^6$ samples are required. In addition, the precision of the estimator $\hat{v}_N$ can be only guaranteed with some confidence as its computation relies on randomized methods. Finally, Theorem 1 only gives a way to estimate $\mathsf{TV}(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_{\tilde{x}})$ for given initial states, whereas (6) requires solving an optimization problem. All of this motivates looking into alternative methods for computing the d_{TV^n} metric.

We start with the case when both ldt-MPs $\mathscr{D}$ and $\tilde{\mathscr{D}}$ have the same state space, say X; let P and $\tilde{P}$ be the corresponding stochastic kernels. Recall that each of them acts as a linear operator on the Banach space $\mathrm{b}\mathscr{B}(X)$, e.g. $Pf(x) = \int_X f(y)P(x, \mathrm{d}y)$. Since $\mathrm{b}\mathscr{B}(X)$ is endowed with a sup-norm, one can define an induced norm on operators, as follows:

$$\|P - \tilde{P}\| := \sup_{f \in \mathrm{b}\mathscr{B}(X), \|f\| \leq 1} \|(P - \tilde{P})f\|.$$

We obtain that $\|P - \tilde{P}\| = \sup_{x \in X} \mathsf{TV}(P(x, \cdot), \tilde{P}(x, \cdot))$ and as a result $\|P - \tilde{P}\| \geq \mathsf{TV}^1(\mathsf{Q}_x, \tilde{\mathsf{Q}}_{\tilde{x}})$. We define a new metric as

$$d_{\mathsf{B}}(\mathscr{D}, \tilde{\mathscr{D}}) := \|P - \tilde{P}\|.$$

The next theorem shows how the latter metric can be used in order to derive upper bounds on d_{TV^n}.

Theorem 2. *For any* $n \in \mathbb{N}_0$ *it holds that*

$$d_{\mathsf{TV}^n}(\mathscr{D}, \tilde{\mathscr{D}}) \leq 2 - 2\left(1 - \frac{1}{2} d_{\mathsf{B}}(\mathscr{D}, \tilde{\mathscr{D}})\right)^n. \tag{11}$$

Proof. From [7, Theorem 2] it follows that

$$\mathsf{TV}(\mathsf{P}^n_x, \tilde{\mathsf{P}}^n_x) \leq 2 - 2\left(1 - \frac{1}{2} d_{\mathsf{B}}(\mathscr{D}, \tilde{\mathscr{D}})\right)^n,$$

for any initial state $x \in X$. Since Q^n_x and $\tilde{\mathsf{Q}}^n_x$ are image measures generated by P^n_x and $\tilde{\mathsf{P}}^n_x$ respectively, it holds that $\mathsf{TV}(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_x) \leq \mathsf{TV}(\mathsf{P}^n_x, \tilde{\mathsf{P}}^n_x)$. Finally, it clearly holds that

$$d_{\mathsf{TV}^n}(\mathscr{D}, \tilde{\mathscr{D}}) \leq \sup_{x \in X} \mathsf{TV}(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_x),$$

[5] Note that for any $i \in [1; N]$, we have that $\mathbf{y}^i$ is a vector of elements of Y of length $n+1$: that is a vector of observations of $\mathscr{D}$ over the time horizon n.

which further leads to (11). □

Before we elaborate on Theorem 2, let us first show an example of how it can be applied to construct finite abstractions with any given precision: that would assure that d_{TV^n} satisfies the third criterion in Section 1. Suppose that we are given an ldt-MP $\mathscr{D} = (X, Y, P, \mathsf{L})$, where Y is a finite alphabet and P is an integral kernel, that is $P(x, \mathrm{d}y) = p(x, y)\mu(\mathrm{d}y)$. Here p is a jointly measurable function and μ is a σ-additive measure on X. We say that p is a density of P with respect to the measure μ: for example if $X = \mathbb{R}^m$ and μ is the Lebesgue measure, p is a common density function. Assume further that the state space X is endowed with some metric d_X compatible with its topology, and is bounded with respect to this metric. Let $(X_i)_{i=1}^m$ be a finite partition of X such that L is constant when restricted to any X_i, and denote by δ_i the diameter of X_i in the metric d_X. Let $x_i \in X_i$ be arbitrary points, and define the finite ldt-MP $\tilde{\mathscr{D}} = (\tilde{X}, \tilde{P}, Y, \tilde{\mathsf{L}})$ as follows: $\tilde{X} = [1; m]$, $\tilde{P}(i, \{j\}) := P(x_i, X_j)$ and $\tilde{\mathsf{L}}(i) = \mathsf{L}(x_i)$. Notice that we cannot compare $\mathscr{D}$ and $\tilde{\mathscr{D}}$ directly using the metric d_{B}: the latter is only defined for ldt-MP that share the same state space. Due to this reason, below we construct an auxiliary ldt-MP $\hat{\mathscr{D}}$ which has an infinite state space X so that $d_{\mathsf{B}}(\mathscr{D}, \hat{\mathscr{D}})$ is well-defined, but whose structure makes it trace-equivalent to the finite ldt-MP $\tilde{\mathscr{D}}$. We require the following Lipschitz-like condition.

Assumption 1. *There exist measurable non-negative functions $\kappa_i : X \to [0, \infty)$, such that $K_i := \int_X \kappa_i(y)\mu(\mathrm{d}y) < \infty$ for all $i \in [1; m]$, and such that*

$$|p(x', y) - p(x'', y)| \le \kappa_i(y) d_X(x', x'') \quad \forall x', x'' \in X_i, \quad \forall y \in X.$$

Theorem 3. *Under Assumption 1 for any $n \in \mathbb{N}_0$ it holds that*

$$d_{\mathsf{TV}^n}(\mathscr{D}, \tilde{\mathscr{D}}) \le 2 - 2\left(1 - \max_{i \in [1;m]} K_i \delta_i\right)^n.$$

Proof. We construct an auxiliary ldt-MP $\hat{D} = (X, \hat{P}, Y, \mathsf{L})$ with a lumpable structure to assure that $d_{\mathsf{TV}^n}(\hat{\mathscr{D}}, \tilde{\mathscr{D}}) = 0$, and to compare $d_{\mathsf{TV}^n}(\mathscr{D}, \hat{\mathscr{D}})$ using Theorem 2. For this purpose we let $\hat{P}(x, \mathrm{d}y) = \hat{p}(x, y)\mu(\mathrm{d}y)$ to be an integral kernel where the density $\hat{p}$ is defined as $\hat{p}(x, y) = p(x_i, y)$ for all points $x \in X_i$ and $y \in X$. For any $f \in \mathrm{b}\mathscr{B}(X)$, $\|f\| \le 1$ and $x \in X_i$ it holds that

$$\begin{aligned} |(P - \tilde{P})f(x)| &= \left| \int_X f(y)(p(x, y) - p(x_i, y))\mu(\mathrm{d}y) \right| \\ &\le \int_X |f(y)| \delta_i \kappa_i(y) \mu(\mathrm{d}y) \le K_i \delta_i, \end{aligned}$$

and hence $\|P - \tilde{P}\| \le \max_{i \in [1;m]} K_i \delta_i$, so that

$$d_{\mathsf{TV}^n}(\mathscr{D}, \hat{\mathscr{D}}) \le 2 - 2\left(1 - \max_{i \in [1;m]} K_i \delta_i\right)^n$$

by Theorem 2. Let $\iota : X \to \tilde{X}$ be the map defined by $\iota(x) = i$ iff $x \in X_i$. For any initial state $x \in X$ it holds that $\mathsf{Q}^n_x = \tilde{\mathsf{Q}}^n_{\iota_* x}$: indeed, these measures coincide on each element of the finite set Y^{n+1} by construction. As a result, $d_{\mathsf{TV}^n}(\hat{\mathscr{D}}, \tilde{\mathscr{D}}) = 0$, which completes the proof of the theorem. □

This theorem shows that in case an original ldt-MP satisfies Assumption 1, for any $\varepsilon > 0$ it is possible to construct a finite ldt-MP that is ε-trace equivalent to the original one in the d_{TV^n} metric. As a result, the metric d_{TV^n} satisfies all the criteria in Section 1. Note also that to estimate the metric d_{TV^n}, which quantifies the difference between path measures, namely the *global* (in time) dynamics, we have used the metric d_{B}, which compares the stochastic kernels, hence measuring the difference between transitions of ldt-MP – their *local* dynamics. The difference in local dynamics is often easier to assess, which in turn provides a method to argue about similarity of the global dynamics – see also the discussion in [16, Section II.B]. To formally speak about the similarity of the local dynamics for the processes with different state spaces, we would need to introduce notions of approximate bisimulations. For example, in such a case we could compare $\mathscr{D}$ and $\tilde{\mathscr{D}}$ in the proof of Theorem 3 directly, without the need to introduce the auxiliary ldt-MP $\hat{\mathscr{D}}$. However the discussion on approximate bisimulation goes beyond the scope of this contribution.

Theorems 2 and 3 further improve the formula-free abstraction procedure for ldt-MP introduced in [31]: the bounds on the path measures there increase linearly in the time horizon, whereas (11) provides bounds which never exceed 2 – the maximal meaningful value of the total variation distance. Moreover, such bounds are tight, that is in some special cases the equality in (11) holds.

3.2 Wasserstein distance

Above we have shown the usefulness of the metric d_{TV^n} for LT model-checking. Although this metric is based on the total variation distance, which has an important characterization through coupling (13), we have not used extensively the notion of coupling in the discussion on d_{TV^n} (though it is important in the proof [7, Theorem 2] used in Theorem 2).

The notion of coupling appears to be much more important for another metric that we consider next. Suppose that we are given two ldt-MP $\mathscr{D} = (X, Y, P, \mathsf{L})$ and $\tilde{\mathscr{D}} = (\tilde{X}, \tilde{P}, Y, \tilde{\mathsf{L}})$ expressed as

$$\begin{cases} \mathbf{x}_{k+1} &= F(\mathbf{x}_k, w_k), \\ \mathbf{y}_k &= \mathsf{L}(\mathbf{x}_k) \end{cases} \quad \begin{cases} \tilde{\mathbf{x}}_{k+1} &= \tilde{F}(\tilde{\mathbf{x}}_k, \tilde{w}_k), \\ \tilde{\mathbf{y}}_k &= \tilde{\mathsf{L}}(\tilde{\mathbf{x}}_k) \end{cases}$$

where each of the $(w_k)_{k \in \mathbb{N}_0}, (\tilde{w}_k)_{k \in \mathbb{N}_0}$ is a sequence of iid random variables. Suppose further that we are interested in approximating $\mathscr{D}$ with $\tilde{\mathscr{D}}$. If the observation space Y is endowed with some metric d_Y, we can endow Y^{n+1} with the product metric

$$d_{Y^{n+1}}((y_0, \dots, y_n), (\tilde{y}_0, \dots, \tilde{y}_n)) := \max_{k \in [0;n]} d_Y(y_k, \tilde{y}_k).$$

If we are able to assure that $d_{Y^{n+1}}(\mathbf{y}, \tilde{\mathbf{y}}) \le \delta$ with a high probability, the fact that $\tilde{\mathbf{y}}$ satisfies some property $\Phi \in \mathrm{b}\mathscr{B}(Y^{n+1})$ would imply that $\mathbf{y}$ satisfies the modified (inflated) property

$$\Phi_\delta = \{y \in Y^{n+1} : d_{Y^{n+1}}(y, \Phi) \le \delta\}.$$

See [19, Theorem 7] for the case when Φ is a reachability specification. However, in order to talk about the probability of the value of $d_{Y^{n+1}}(\mathbf{y}, \tilde{\mathbf{y}})$ being less than δ, we need to consider a coupling between $\mathbf{y}$ and $\tilde{\mathbf{y}}$. The work in [19] has considered the case of linear jump-diffusions, and implicitly suggested to use a coupling that matches the noises, that is $\tilde{w}_k = w_k$. In the follow-up paper [2] the coupling was considered to be the independent one (cf. Example 4). However, one has a complete freedom in choosing the coupling, so we can define a metric as follows:

$$d_{\mathsf{W}^n}(\mathscr{D}, \tilde{\mathscr{D}}) := \sup_{x \in X} \inf_{\tilde{x} \in \tilde{X}} \mathsf{W}(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_{\tilde{x}}),$$

thanks to the coupling characterization of the Wasserstein distance in (14). In case we know $d_{\mathsf{W}^n}(\mathscr{D}, \tilde{\mathscr{D}})$, for any initial state $x \in X$ we can find a corresponding $\tilde{x} \in \tilde{X}$ such that

$$\mathbb{Q}\left(d_{Y^{n+1}}(\mathbf{y}, \tilde{\mathbf{y}}) \ge \delta\right) \le \frac{d_{\mathsf{W}^n}(\mathscr{D}, \tilde{\mathscr{D}})}{\delta},$$

thanks to Markov's inequality, where $\mathbb{Q} \in \Gamma(\mathsf{Q}^n_x, \tilde{\mathsf{Q}}^n_{\tilde{x}})$ is any among the optimal couplings in (14). As a result, the metric d_{W^n} as well satisfies the first criterion in Section 1. Randomized methods to compute this metric can be found in [14, Sections 2.2, 2.3] and [28], thus assuring that d_{W^n} satisfies the second criterion we have raised. These methods are based on the representation of W as an integral metric (see Section 7.2), for example the one in [28] suggests solving a linear programming problem to evaluate $\mathsf{W}(\mathsf{Q}^{n,N}_x, \tilde{\mathsf{Q}}^{n,N}_{\tilde{x}})$ over empirical distributions (8).

Both TV and W are integral metrics, and for bounded metric spaces TV is always bigger than W, multiplied by a constant. Hence, one may expect that d_{W^n} provides less conservative estimates than d_{TV^n}. Unfortunately, although one can try defining an analogue of the transition-based metric d_{B} for d_{W^n}, obtaining estimates similar to Theorem 2 is not an easy task. Thus we are not able to claim that d_{W^n} satisfies the third criterion.

Let us also mention that the Wasserstein metric has also been applied for purposes of abstractions of controlled ldt-MP in [14] with focus on discounted additive cost functionals. Furthermore, [30] has employed the Wasserstein metric to compare outputs of stochastic processes. However, the results in both papers are not suitable for Problem 1.

4. COMPARISON WITH THE LITERATURE

4.1 Total Variation distance

Recent literature on approximate probabilistic bisimulations has led to several metrics for discrete-time stochastic systems that are related to the metrics d_{B} and d_{TV^n} considered here. Before we elaborate on their similarity, let us first recall a few of these metrics.

One of the first metrics for stochastic systems is introduced in [11] and [12] for finite and infinite systems, respectively. As we discussed in Section 1, the notion of exact probabilistic bisimulation, introduced for finite systems in [22] and extended to infinite systems in [10], appears to be too restrictive and to lack robustness: this motivates looking for more flexible relations between systems. The notion of exact probabilistic bisimulation is characterized via a binary logic $\mathscr{L}$, so that two states are bisimilar if and only if they satisfy the same formulae in $\mathscr{L}$. The extension of $\mathscr{L}$ to some *real-valued* logic $\mathscr{L}_{\mathrm{r}}$ leads to the definition of a distance between two states x and $\tilde{x}$ as the maximal difference between all formulae of $\mathscr{L}_{\mathrm{r}}$ computed over such states: $d_{\mathscr{L}_{\mathrm{r}}}(x,\tilde{x}) := \sup_{f\in\mathscr{L}_{\mathrm{r}}} |f(x) - f(\tilde{x})|$. The fact that $\mathscr{L}_{\mathrm{r}}$ is an extension of $\mathscr{L}$ requires that $d_{\mathscr{L}_{\mathrm{r}}}$ is equal to zero over bisimilar states. The value of the metric between two stochastic systems is defined (over the disjoint union of the two systems) as the $d_{\mathscr{L}_{\mathrm{r}}}$ distance between their initial conditions.

A different approach is taken in [35], where stochastic systems are represented as co-algebras, namely pairs (X, P), where X is a state space and $P : X \to \mathscr{P}(X)$ (note the similarity with dt-MP models considered in this work). The state spaces are assumed to be endowed with a metric, and the functor $\mathscr{P}$ is assumed to push the original metric of X into a Kantorovich metric over $\mathscr{P}(X)$.[6] It is further shown that a final co-algebra exists for such a functor, and since the final co-algebra is often thought of as a collection of behaviors, [35] introduces a metric between two states $x, \tilde{x} \in X$ as $d_{\mathrm{co}}(x,\tilde{x}) = d_{\mathrm{fin}}(!(x), !(\tilde{x}))$, where ! is the unique map from X to the final co-algebra.

Although $d_{\mathscr{L}_{\mathrm{r}}}$ and d_{co} are introduced in completely different ways, it is shown in [35] that under certain conditions $d_{\mathrm{co}} = d_{\mathscr{L}_{\mathrm{r}}}$. This could be considered as evidence that any of these two (equal) metrics is the natural metric for the intended applications. However, it is of interest whether either of the two approaches is *the* natural one. Indeed, the definition of $d_{\mathscr{L}_{\mathrm{r}}}$ depends on the choice of the real-valued logic $\mathscr{L}_{\mathrm{r}}$, which in [12] is chosen to be just one of many possible extensions of the binary logic $\mathscr{L}$. Similarly, given a metric on the Borel space X, there are many ways to introduce a related metric on $\mathscr{P}(X)$: besides the Kantorovich metric proposed in [35] one could employ a Prokhorov metric [15] which would lead to a different value for d_{co}. As convergence in Prokhorov metric is equivalent to weak convergence of measures, and hence to the topology of $\mathscr{P}(X)$, arguably the latter may be a more natural choice for a metric on $\mathscr{P}(X)$ than the Kantorovich one. In conclusion, we contend that it is in general not possible to assert how good or natural the choice of a particular metric is, based on the way this metric is defined: the usefulness of a metric is rather determined by the applications it is meant to be used for. From this more practical perspective, it is unfortunately not clear which applications the models studied in [12, 35] are suitable for, besides their use in testing [34]: in fact, their semantics has not been defined explicitly and constructively. Besides, up to the best of the authors knowledge, in the case of testing there exist no precise bounds on the difference between testing probabilities over two system given their d_{co} or $d_{\mathscr{L}_{\mathrm{r}}}$ distances.

From the perspective of the preceding discussion, we aim next to elaborate on the similarity between $d_{\mathscr{L}_{\mathrm{r}}}$ and d_{TV^n}. Let us in particular show how $d_{\mathscr{L}_{\mathrm{r}}}$ can be applied on BLTL model-checking over ldt-MP (which are the models we have defined d_{TV^n} over). Since $d_{\mathscr{L}_{\mathrm{r}}}$ has been introduced over Labelled Markov Processes (LMP) [10], we need to provide a way of transform an ldt-MP into a corresponding LMP[7]. Consider a ldt-MP $\mathscr{D} = (X, Y, P, \mathsf{L})$ with a finite observation space Y; we define a corresponding LMP $(X, Y, \{\tau_y\}_{y\in Y})$, endowed with a state space X, a labels set Y, and the transition kernel

$$\tau_y(x, \mathrm{d}x') := 1_{\mathsf{L}^{-1}(y)}(x) \cdot P(x, \mathrm{d}x'),$$

that is, we enable the label y in the LMP exactly in those states x that are labeled with y in $\mathscr{D}$. Suppose we would like to compute over $\mathscr{D}$ a probability of some basic BLTL formula, that is a finite word $y_0 \dots y_n \in Y^{n+1}$: i.e. we are to find $\mathsf{Q}^n_x(y_0 \dots y_n)$, where Q^n_x is the trace distribution $\mathscr{D}$ induces on Y^{n+1}. As LMP do not have explicitly-defined trace semantics, we can interpret $\mathsf{Q}^n_x(y_0 \dots y_n)$ as a value of the $\mathscr{L}_r$ formula $\langle y_0\rangle \dots \langle y_n\rangle 1(x)$.[8] The latter interpretation tells us that the LMP can be applied to compute probabilities of basic BLTL formulae over a corresponding ldt-MP, and hence it can be applied to any BLTL formula [31, Section 2.4]. Note however, that probabilities of more general BLTL formulae over ldt-MP are not necessarily elements of $\mathscr{L}_{\mathrm{r}}$ for the corresponding LMP: the disjunction and conjunction in BLTL are not related to max and min in $\mathscr{L}_{\mathrm{r}}$ at all. As a result, although any BLTL formula over an ldt-MP can be computed via functions in $\mathscr{L}_{\mathrm{r}}$ over the corresponding LMP, such computations may require summation of $\mathscr{L}_{\mathrm{r}}$ functions. Hence, if any function in $\mathscr{L}_{\mathrm{r}}$ is known only with some error, this operation will require accumulating errors – this issue was already discussed in [31, Section 3] concerning the possible extensions of safety-focused

[6] Note however, that despite a strong connection between Kantorovich and Wasserstein metrics (see e.g. [14, Section 2.4.6]), this approach differs from the one we apply here in Section 3.2.

[7] Up to our knowledge, this paper is the first to explicitly elaborate on such a transformation.

[8] We assume here a version of $\mathscr{L}_r$ without discounting.

results in [4] to the whole BLTL. Thus, although [12, Proposition 7.5] concerning $d_{\mathscr{L}_\tau}$ provides an analysis similar to Theorem 2, the bounds from the former result when interpreted over ldt-MP only apply to basic BLTL formulae, whereas the latter result provides bounds for any BLTL formula.

Let us also mention that Theorems 2 and 3 concerning the d_{TV^n} metric improve formula-free guarantees for ldt-MP introduced in [31] by providing tighter bounds on the difference between trace distributions: indeed, [31, Lemma 1, Theorem 3] derives similar bounds as Theorem 2, but provide more conservative results. An alternative approach [5] suggests expressing an LT property as an automaton, and solving a safety problem over the product system. Unfortunately, the overall error associated to the abstraction needed for the solution of the safety problem depends on the size of the automaton, which is particularly crucial in case of BLTL formulae for which automata can be large [31, Section 3.4]. The formula-free guarantees do not depend on a particular BLTL formula (as the name suggests), and provide arguably less conservative bounds than safety-based approximate model-checking of complex LT properties (cf. [31, Section 5]).

4.2 Wasserstein distance

As we have mentioned in Section 3.2, the definition of the d_{W^n} metric is inspired by the work [19], which introduces an analogous metric restricted to a fixed coupling structure. We contend that this feature substantially increases the conservatism of the results, since in general one has the freedom to optimize over the coupling. The notions in [19] hinge upon similar ones developed over non-stochastic systems [16], where the coupling is necessarily unique so that its choice does not play a role. The choice of coupling for stochastic processes is especially important when it is not meaningful to claim that both processes are driven by the same noise. For instance, if it is clear how to compare noises that drive two diffusions (cf. Example 4), it is much less clear how to couple a diffusion with a finite-space continuous-time Markov Chain serving as its approximation [2]: in the latter case the only natural coupling seems to be the independent one, and it is rather unlikely that this coupling is the optimal one.

The work in [19] suggests how to compute a metric over a fixed coupling under rather strong assumptions on the stability of the models (over a fixed coupling). In particular, these technique require Lyapunov-like functions to synthesize a metric, and provide practically relevant sufficient conditions for these functions only in the linear case. In contrast, [6] introduces randomized methods to compute fixed coupling metrics without any stability assumptions – as in our case the results are valid over a finite time horizon. It is of interest to compare those results with ours, to see whether the choice of the coupling leads to a reduced conservatism, as expected.

Let us consider an example drawn form [1] and also studied in [32]. We consider a regional power network consisting of two local subnetworks: each of them has its own energy storage capacity. There are two sources of the energy: a coal plant shared by both networks, and two separate wind farms producing renewable energy. [32] studied the approximate abstraction of this model, with the objective of coal plant energy production and distribution optimizing some desired property over the whole network. On the other hand, here we are interested in studying the effect of the correlation over the energy produced by the two wind farms – it practically makes sense to consider this correlation when the two local subnetworks are close geographically. We assume that the coal plant energy is evenly distributed between both networks, which leads to the following (autonomous) dt-MP model of the network:

$$\begin{cases} \mathbf{x}^1_{k+1} = \beta\left(\mathbf{x}^1_k + \frac{1}{2}p_k + r^1_k - d^1_k\right) \vee 0 \wedge M, \\ \mathbf{x}^2_{k+1} = \beta\left(\mathbf{x}^2_k + \frac{1}{2}p_k + r^2_k - d^2_k\right) \vee 0 \wedge M, \end{cases}$$

where $\mathbf{x}^i$ it the amount of the stored energy in the i-th subnetwork, $\beta = 0.8$ is the loss rate of the stored energy, p_k is the energy produced by the coal plant, r^i_k is a renewable energy produced by the i-th wind farm, d^i_k is the energy demand in the i-th network, and $M = 30$ is the max storage capacity. We assume that $p_k \sim U([3,5])$ and $d^i_k \sim U([1,2])$ are distributed uniformly. The distribution of the renewable energy instead consists of two components: $r^i_k = (1-\rho)e^i_k + \rho r_k$, where $e^i_k, r_k \in \mathscr{E}(1)$ are distributed exponentially and $\rho \in [0,1]$ measures the correlation level of the distributions r^1_k and r^2_k. The initial conditions $\mathbf{x}^i_0 \sim U([5,10])$ are distributed uniformly.

We study two copies of the model distinguished by the choice of parameter $\rho = 0.2$ and $\rho = 0$ respectively. Recall that we only know the distribution of the stochastic process associated to each of the two models, but not their joint distribution: this allows to choose the coupling. We then consider three possible choices: coupling via the same noise, independence coupling, and reflected coupling. In the former case we use the same samples of $\tilde{d}^i_k = d^i_k$, $\tilde{p}_k = p_k$ and $\tilde{\mathbf{x}}^i_0 = \mathbf{x}^i_0$ for both systems, whereas in the latter case we use the symmetry of the uniform distribution to define $\tilde{d}^i_k$, $\tilde{p}_k$ and $\tilde{\mathbf{x}}^i_0$ as d^i_k, p_k and $\mathbf{x}^i_0$ reflect with respect to their mean values. Finally, in case of the independence coupling all samples are assumed to be independent, that is $\tilde{d}^i_k \perp d^i_k$, $\tilde{p}_k \perp p_k$ and $\tilde{\mathbf{x}}^i_0 \perp \mathbf{x}^i_0$. We assume that the observation space corresponds to the state space endowed with the Euclidean metric, thus the observation map is the identity function.

The results are presented in Figure 1. The computations are run over a time horizon $n = 100$ for $N \in [2;1000]$ samples. From the figure it can be seen that over the number of samples considered, for each estimator the convergence does hold. One can also observe that although the coupling via the same noise performs better than the independence coupling, or than that via reflected noise, their associated distance is greater than the distance $d_{\mathsf{W}^{100}}$ obtained by optimizing over all the possible couplings. This optimization is performed using the method in [28], by means of solving a linear programming problem: the latter clearly represents the computational bottleneck of the method, as it requires $2N$ variables and $N(N-1)$ constraints. The complexity then depends on the number of samples, and although there is no explicit connection between the complexity of the linear programming problem and the dimension of original systems (or the time horizon), the latter factor may affect the number of samples needed for convergence to the true solution. Unfortunately, up to our knowledge there are no explicit results on the convergence for the W metric, similar to Theorem 1 for the TV metric. Methods mentioned in [14, Sections 2.2, 2.3] can be explored as possible computationally efficient alternatives. Let us finally mention that we do not provide here a comparison over the d_{TV^n} metric as it would not enlighten the importance of coupling as significantly as the d_{W^n} metric does.

5. CONCLUSIONS

This paper has discussed the formulation of two new metrics for stochastic transition systems, with focus on their applications for linear-time model-checking. Such metrics are shown to be useful according to three criteria defined in the article, and are

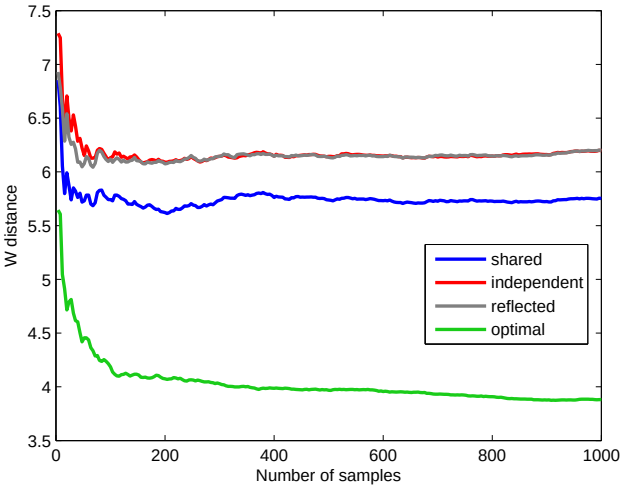

Figure 1: Evaluation of expected distances between x and $\tilde{x}$ depending on the number of samples, over different couplings: shared (blue line in the middle), independent and reflected noises (resp. red and gray lines on the top), optimal coupling (green line at the bottom).

claimed to perform better compared to other metrics proposed in the literature. Table 1 further shows which of those criteria the metrics introduced in this paper satisfy, and according to which result. A promising direction of future research concerns the connections of the metrics proposed in this work with notions of approximate probabilistic simulation and bisimulation.

Criterion	TV^n	W^n
[i.]	YES. By definition.	YES. By definition.
[ii.]	YES. By Theorem 1.	YES. See e.g. [28].
[iii.]	YES. By Theorems 2, 3.	???

Table 1: Satisfaction of criteria by metrics TV^n and W^n.

6. REFERENCES

[1] MoVeS website. http://www.movesproject.eu.

[2] A. Abate. A contractivity approach for probabilistic bisimulations of diffusion processes. In *Proceedings of the 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference*, pages 2230–2235, 2009.

[3] A. Abate. Approximation metrics based on probabilistic bisimulations for general state-space markov processes: a survey. *Electronic Notes in Theoretical Computer Sciences*, 2013. In Print.

[4] A. Abate, J.-P. Katoen, J. Lygeros, and M. Prandini. Approximate model checking of stochastic hybrid systems. *European Journal of Control*, 16:624–641, 2010.

[5] A. Abate, J.-P. Katoen, and A. Mereacre. Quantitative automata model checking of autonomous stochastic hybrid systems. In *Proceedings of the 14th international conference on Hybrid Systems: Computation and Control*, HSCC '11, pages 83–92, New York, NY, USA, 2011. ACM.

[6] A. Abate and M. Prandini. Approximate abstractions of stochastic systems: A randomized method. In *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference*, pages 4861–4866, 2011.

[7] A. Abate, F. Redig, and I. Tkachev. On the effect of perturbation of conditional probabilities in total variation. *Statistics & Probability Letters*, 2014. In Press.

[8] C. Baier and J.-P. Katoen. *Principles of model checking*. The MIT Press, 2008.

[9] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer-Verlag, Berlin, 2010.

[10] J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for Labelled Markov Processes. *Information and Computation*, 179(2):163 – 193, 2002.

[11] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Metrics for Labeled Markov Systems. In *Proceedings of the 10th International Conference on Concurrency Theory*, CONCUR '99, pages 258–273, London, UK, 1999. Springer-Verlag.

[12] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Metrics for labelled Markov processes. *Theoretical Computer Science*, 318(3):323–354, 2004.

[13] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.

[14] N. Ferns, P. Panangaden, and D. Precup. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

[15] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[16] A. Girard and G. J. Pappas. Approximation metrics for discrete and continuous systems. *IEEE Transactions on Automatic Control*, 52(5):782 –798, 2007.

[17] A. Girard and G. J. Pappas. Approximate bisimulation: A bridge between computer science and control theory. *European Journal of Control*, 17(5-6):568–578, 2011.

[18] A. Hinton, M. Kwiatkowska, G. Norman, and D. Parker. PRISM: A tool for automatic verification of probabilistic systems. In H. Hermanns and J. Palsberg, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, volume 3920 of *Lecture Notes in Computer Science*, pages 441–444. Springer Verlag, 2006.

[19] A.A. Julius and G.J. Pappas. Approximations of stochastic hybrid systems. *IEEE Transactions on Automatic Control*, 54(6):1193–1203, 2009.

[20] O. Kallenberg. *Foundations of modern probability*. Probability and its Applications. Springer-Verlag, New York, 1997.

[21] J.-P. Katoen, M. Khattri, and I. S. Zapreev. A Markov reward model checker. In *Proceedings of the Second International Conference on the Quantitative Evaluation of Systems*, QEST '05, pages 243–244, Washington, DC, USA, 2005. IEEE Computer Society.

[22] K. G. Larsen and A. Skou. Bisimulation through probabilistic testing. *Information and Computation*, 94(1):1–28, 1991.

[23] T. Lindvall. *Lectures on the coupling method*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1992.

[24] B. Øksendal. *Stochastic differential equations: an introduction with applications*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003.

[25] K. R. Parthasarathy. *Probability measures on metric spaces*. Probability and Mathematical Statistics, No. 3. Academic Press Inc., New York, 1967.

[26] D. Revuz. *Markov chains*. North-Holland Publishing, Amsterdam, second edition, 1984.

[27] R. Segala. A compositional trace-based semantics for probabilistic automata. In I. Lee and S. A. Smolka, editors, *CONCUR '95: Concurrency Theory*, volume 962 of *Lecture Notes in Computer Science*, pages 234–248. Springer Berlin Heidelberg, 1995.

[28] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[29] P. Tabuada. *Verification and control of hybrid systems: A symbolic approach*. Springer Verlag, New York, 2009.

[30] D. Thorsley and E. Klavins. Approximating stochastic biochemical processes with Wasserstein pseudometrics. *Systems Biology, IET*, 4(3):193 –211, May 2010.

[31] I. Tkachev and A. Abate. Formula-free finite abstractions for linear temporal verification of stochastic hybrid systems. In *Proceedings of the 16th international conference on Hybrid Systems: Computation and Control*, HSCC '13, pages 283–292, New York, NY, USA, 2013. ACM.

[32] I. Tkachev, A. Mereacre, J.-P. Katoen, and A. Abate. Quantitative automata-based controller synthesis for non-autonomous stochastic hybrid systems. In *Proceedings of the 16th international conference on Hybrid Systems: Computation and Control*, HSCC '13, pages 293–302, New York, NY, USA, 2013. ACM.

[33] Ilya Tkachev and Alessandro Abate. Characterization and computation of infinite-horizon specifications over markov processes. *Theoretical Computer Science*, 515(0):1 – 18, 2014.

[34] F. van Breugel, S. Shalit, and J. Worrell. Testing labelled Markov processes. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, ICALP '02, pages 537–548, London, UK, 2002. Springer-Verlag.

[35] F. van Breugel and J. Worrell. Towards quantitative verification of probabilistic transition systems. In *Proceedings of the 28th International Colloquium on Automata, Languages and Programming*, ICALP '01, pages 421–432, London, UK, 2001. Springer-Verlag.

[36] M.Y. Vardi. Automatic verification of probabilistic concurrent finite state programs. In *26th Annual Symposium on Foundations of Computer Science*, pages 327–338, 1985.

7. APPENDIX

7.1 General notation

We use extensively standard notions from measure and probability theory, for precise definitions see e.g. [20]. A measurable space is a pair $(E,\mathscr{E})$ where E is an arbitrary set and $\mathscr{E}$ is a σ-algebra on E. The space of all probability measures on $(E,\mathscr{E})$ is denoted by $\mathscr{P}(E,\mathscr{E})$; it is assumed to be endowed with the smallest σ-algebra that makes any evaluation map $\theta_A : \mathscr{P}(E,\mathscr{E}) \to \mathbb{R}$, defined by $\theta_A(p) = p(A)$ for all $p \in \mathscr{P}(E,\mathscr{E})$ and $A \in \mathscr{E}$, measurable. A stochastic kernel on $(E,\mathscr{E})$ is a measurable map $K : E \to \mathscr{P}(E,\mathscr{E})$. A probability space is a triple $(E,\mathscr{E},p)$, where $(E,\mathscr{E})$ is a measurable space and $p \in \mathscr{P}(E,\mathscr{E})$. If $(\Omega,\mathscr{F},\mathsf{P})$ is a probability space and $f : (\Omega,\mathscr{F}) \to (E,\mathscr{E})$ is a measurable map then

$$(f_*\mathsf{P})(A) := \mathsf{P}\left(f^{-1}(A)\right), \quad \forall A \in \mathscr{E}$$

defines the probability measure in $\mathscr{P}(E,\mathscr{E})$. We say that f induces the measure $f_*\mathsf{P}$. We further denote by id_Ω the identity map on Ω; clearly, it holds that $(\mathrm{id}_\Omega)_*\mathsf{P} = \mathsf{P}$ for any $\mathsf{P} \in \mathscr{P}(\Omega,\mathscr{F})$. By $\otimes$ we denote the product of measures or σ-algebras. The set of real numbers is denote by $\mathbb{R}$, the set of natural numbers by $\mathbb{N}$ and we write $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We write E^ω in place of $E^{\mathbb{N}_0}$.

Any topological space X is assumed to be endowed with its Borel σ-algebra $\mathscr{B}(X)$ generated by all open subsets of X. For topological spaces we simply write $\mathscr{P}(X)$ in place of $\mathscr{P}(X,\mathscr{B}(X))$. We refer to elements of $\mathscr{B}(X)$ as Borel subsets of X. A (standard) Borel space is a topological space that is homeomorphic to a Borel subset of complete separable metric space. By $\mathscr{C}(X)$ we denote the linear space of all continuous real-valued functions on X.

Given a measurable space $(E,\mathscr{E})$, we denote by $\mathrm{b}\mathscr{E}$ the linear space of all bounded functions $f : E \to \mathbb{R}$. It is a Banach space endowed with a sup-norm $\|f\| := \sup_{x\in E} |f(x)|$.

7.2 Distances between probability measures

For a brief overview on distance between probability measures see [15]; a detailed exposition can be found in [25]. In this paper we focus on two metrics: the total variation distance TV that applies to measures over any measurable spaces, and the Wasserstein distance W that requires the underlying space to be a metric space. Both metrics fall into the class of integral probability metrics: let $(E,\mathscr{E})$ be an some measurable space and let $\mathscr{G}$ be any collection real-valued measurable functions on E. The corresponding metric on $\mathscr{P}(E,\mathscr{E})$ is defined as

$$I_{\mathscr{G}}(\mu,\tilde{\mu}) := \sup_{g\in\mathscr{G}} \left| \int_E g \,\mathrm{d}(\mu - \tilde{\mu}) \right|.$$

The total variation metric is given by $\mathsf{TV} = I_{\mathrm{b}\mathscr{E}_1}$ where $g \in \mathrm{b}\mathscr{E}_1$ iff $g \in \mathrm{b}\mathscr{E}$ and $\|g\| \leq 1$. Alternatively, the metric TV can be introduced via the difference of μ and $\tilde{\mu}$ over sets as follows:

$$\mathsf{TV}(\mu,\tilde{\mu}) = 2 \cdot \sup_{A\in\mathscr{E}} |\mu(A) - \tilde{\mu}(A)|. \tag{12}$$

Unlike the total variation distance that depends exclusively on the measurability structure of the set E, the Wasserstein distance requires E to be a metric space. Let E be endowed with the metric d_E such that $\mathscr{E} = \mathscr{B}(E)$. For any $g \in \mathrm{b}\mathscr{E}$ let

$$\mathrm{Lip}(g) := \sup_{x\neq x'} \frac{|g(x) - g(x')|}{d_E(x,x')}$$

denote the Lipschitz constant of g. The Wasserstein distance is defined as $\mathsf{W} = I_{\mathscr{C}_1}$, where $g \in \mathscr{C}_1$ iff $\mathrm{Lip}(g) \leq 1$. Provided that $\mathrm{diam}(E) := \sup_{x,x'\in E} d_E(x,x') < \infty$, the following inequality relates the two metrics: $\mathsf{TV} \geq \frac{1}{\mathrm{diam}(E)}\mathsf{W}$ [15], so that the convergence in TV implies the convergence in W in case the latter is defined using a metric in which E is bounded.

Another relation between TV and W can be derived via recalling that each of the metric can be characterized by a coupling, optimal in a certain sense. For $\mu,\tilde{\mu} \in \mathscr{P}(E,\mathscr{E})$ let us denote by

$\Gamma(\mu,\tilde{\mu}) \subset \mathscr{P}(E^2,\mathscr{E}^2)$ the collection of all coupling measures for μ and $\tilde{\mu}$. Provided that E is a Borel space and $\mathscr{E} = \mathscr{B}(E)$:

$$\mathsf{TV}(\mu,\tilde{\mu}) = 2 \cdot \sup_{M \in \Gamma(\mu,\tilde{\mu})} M(\Delta_E) \tag{13}$$

and if (E, d_E) is a bounded complete separable metric space:

$$\mathsf{W}(\mu,\tilde{\mu}) = \inf_{M \in \Gamma(\mu,\tilde{\mu})} \int_{E \times E} d_E(x,\tilde{x}) M(\mathrm{d}x \times \mathrm{d}\tilde{x}). \tag{14}$$

If we consider any two random elements f and $\tilde{f}$ that represent distributions μ and $\tilde{\mu}$, then the total variation distance can be obtained via the coupling that maximizes the probability that $f = \tilde{f}$, whereas the Wasserstein distance is exactly the minimal expected value of $d_E(f,\tilde{f})$ over all possible couplings.