BISIMULATION METRICS FOR CONTINUOUS MARKOV DECISION PROCESSES

NORM FERNS[*], PRAKASH PANANGADEN[†], AND DOINA PRECUP[‡]

Abstract. In recent years, various metrics have been developed for measuring the behavioural similarity of states in probabilistic transition systems [Desharnais et al., Proceedings of CONCUR, (1999), pp. 258-273, van Breugel and Worrell, Proceedings of ICALP, (2001), pp. 421-432]. In the context of finite Markov decision processes, we have built on these metrics to provide a robust quantitative analogue of stochastic bisimulation [Ferns et al., Proceedings of UAI, (2004), pp. 162-169] and an efficient algorithm for its calculation [Ferns et al., Proceedings of UAI (2006), pp.174-181]. In this paper, we seek to properly extend these bisimulation metrics to Markov decision processes with continuous state spaces. In particular, we provide the first distance-estimation scheme for metrics based on bisimulation for continuous probabilistic transition systems. Our work, based on statistical sampling and infinite dimensional linear programming is a crucial first step in formally guiding real-world planning, where tasks are usually continuous and highly stochastic in nature, e.g. robot navigation, and often a substitution with a parametric model or crude finite approximation must be made. We show that the optimal value function associated with a discounted infinite-horizon planning task is continuous with respect to metric distances. Thus, our metrics allow one to reason about the quality of solution obtained by replacing one model with another. Alternatively, they may potentially be used directly for state aggregation. An earlier version of this work appears in the doctoral thesis of Norm Ferns [McGill University, (2008)].

Key words. bisimulation, metrics, reinforcement learning, continuous, Markov decision process

AMS subject classifications. 90C40, 93E20, 68T37, 60J05

1. Introduction.

Markov decision processes (MDPs) offer a popular mathematical tool for planning and learning in the presence of uncertainty [7]. They are a standard formalism for describing multi-stage decision making in probabilistic environments where the objective of the decision making is to maximize a cumulative measure of long-term performance, called the *return*. Dynamic programming algorithms, e.g., value iteration, policy iteration [53], allow one to compute the optimal expected return for any state, as well as the way of behaving, or policy, that generates this return. However, in many practical situations the state space of an MDP may be too large, possibly continuous, for the standard algorithms to apply. Similarly, MDPs with a high degree of stochasticity, i.e., when there are many possible outcome states for probabilistic state transitions, can be much more problematic to solve than those that are nearly deterministic [43]. Therefore, one usually turns to model approximation to find a simpler relevant model. The hope is that this can be done in such a manner so as to construct an "essentially equivalent" MDP with significantly reduced complexity, thereby allowing the use of classical solution methods while at the same time providing a guarantee that solutions to the reduced MDP can be extended to the original.

Recent MDP research on defining equivalence relations on MDPs [11, 32] has built on the notion of strong probabilistic bisimulation from concurrency theory. Probabilistic bisimulation was introduced by [41] based on bisimulation for nondeterministic systems due to [50] and [44]. Henceforth when we say "bisimulation" we will mean strong probabilistic bisimulation.

In a probabilistic setting, bisimulation can be described as an equivalence relation that relates two states precisely when they have the same probability of transitioning to classes of equivalent states. The extension of bisimulation to transition systems with rewards was carried out in the

[*] This work was supported by the FQRNT and by NSERC.

[†] Supported by NSERC and by the Office of Naval Research

[‡] Supported by NSERC and the Office of Naval Research.

context of MDPs by [32] and in the context of performance evaluation by [3]. In both cases, the motivation is to use the equivalence relation to aggregate the states and get smaller state spaces. The basic notion of bisimulation is modified only slightly by the introduction of rewards.

However, it has been well established that the use of exact equivalences in quantitative systems is problematic. A notion of equivalence is two-valued: two states are either equivalent or they are not. For example, a small perturbation of the transition probabilities of a probabilistic transition system can make two equivalent states no longer equivalent. In short, any kind of equivalence is unstable - too sensitive to perturbations in the numerical values of the parameters of the model.

A natural remedy is to use pseudometrics. A pseudometric is almost the same as a metric, except that two distinct points can be at zero distance. Given a pseudometric, we define an equivalence relation by saying that two points are equivalent if they are at zero distance; this is called the *kernel* of the pseudometric. We will just say "metric" henceforth. Metrics are natural quantitative analogues of equivalence relations. The triangle inequality, for example, can be interpreted as a quantitative generalization of transitivity: if states x_1 and x_2, and x_2 and x_3, are close in distance then so too must be states x_1 and x_3. The metrics on which we focus here specify the degree to which objects of interest behave similarly; usually we would like the kernel to be bisimilarity, the largest bisimulation relation.

Much of this work has been done in a very general setting, using the labelled Markov process (LMP) model [5, 15, 49]. Previously defined metrics [16, 59, 18, 17] are quantitative generalizations of bisimulation; they assign distance zero to states that are bisimilar, distance one to states that are easily distinguishable, and an intermediate distance to those in between.

Van Breugel and Worrell (2001) [59] showed how, in a simplified setting of finite state space LMPs, metric distances could be calculated in polynomial time. This work, along with that of others [18], was then adapted to finite MDPs [27]. The current authors used fixed-point theory to construct metrics, each of which had bisimilarity as its kernel, was sensitive to perturbations in MDP parameters, and provided bounds on the optimal values of states. We showed how to compute the metrics up to any prescribed degree of accuracy and then used them to directly aggregate sample finite MDPs. We subsequently discovered a more efficient method for estimating metrics based on statistical sampling and network optimization [26].

In this paper, we present a significant generalization of these previous results to MDPs with continuous state spaces. The linear programming arguments we used in our previous work no longer apply, and we have to use measure theory and duality theory on continuous state spaces. The mathematical theory is interesting in its own right. Although continuous MDPs are of great interest for practical applications, e.g. in the areas of automated control and robotics, the existing methods for measuring distances between states, for the purpose of state aggregation as well as other approximation methods are still largely heuristic. As a result, it is hard to provide guaranteed error bounds between the correct and the approximate value function. It is also difficult to determine the impact that structural changes in the approximation technique would have on the quality on the approximation. The metrics we define in this paper allow the definition of error bounds for value functions. These bounds can be used as a tool in the analysis of existing approximation schemes.

An earlier version of this work appears in [25]. The existence of the metrics and some continuity results in a continuous setting were originally presented in less polished form in [28]; here we unify and strengthen those results. Specifically, the main contributions of this work are:

(i) We extend an approach to bisimulation metrics for finite state probabilistic transition systems due to [59], based on linear programming, to bisimulation metrics for continuous state space Markov decision processes using infinite dimensional linear programming (Theorem 3.12).

This is a refinement of previous work [28].

(ii) We prove Lipschitz continuity of the optimal value function with respect to our bisimulation metrics for continuous state space Markov decision processes (Theorem 3.20). This is a refinement of previous work [28].

(iii) Our key result is to extend the metric approximation scheme, developed in [26] for finite MDPs, to a continuous setting (compact metric spaces).

The rest of the paper is organized as follows: in § 2, we present a review of the theory of finite Markov decision processes as it pertains to the standard reinforcement learning paradigm, bisimulation, and bisimulation metrics. We also provide a brief survey of mathematics for continuous spaces to set down the notations and results relevant for subsequent sections. Section 3 shifts the discussion to Markov decision processes with infinite state spaces, introducing issues of measurability and continuous analogues of concepts introduced in § 2. We use properties of the Kantorovich functional, an infinite linear program that can be used to define a metric on probability measures, to arrive at our first major result: existence of bisimulation metrics, along with several continuity properties. We establish an important reinforcement-learning bound and a simple calculation, illustrating the use of metric reasoning. In § 4 we provide a brief mathematical background of empirical processes, including a crucial Glivenko-Cantelli theorem. In § 5 and § 6we then present our central result: an approximation scheme for estimating distances for MDPs whose state spaces are compact metric spaces. We attempt to bound the running time and estimation error of this approximation scheme in § 7. Finally, in § 8 we conclude with a summary of our results, related work, and directions for further research.

2. Background.

In this section we first review the basics of finite Markov decision processes with respect to reinforcement learning, bisimulation, and bisimulation metrics. We assume the reader is familiar with basic discrete mathematics, including discrete probability theory and finite metric spaces. Next we set down in some detail fundamental mathematical results for continuous spaces relevant for subsequent sections. Some of the issues that arise there are quite subtle; thus, we clearly set down the notation and results to be used to avoid any ambiguity.

2.1. Reinforcement Learning.

We define reinforcement learning to be that branch of artificial intelligence that deals with an agent learning through interaction with its environment in order to achieve a goal. The intuition behind reinforcement learning is that of learning by trial and error. By contrast, in supervised learning an external supervisor provides examples of desired behaviour from which an agent can learn, much as a student learns from a teacher.

Applications of reinforcement learning include optimal control in robotics [40], meal provisioning [34], scheduling, brain modelling, game playing, and more.

The interaction of an agent with its environment in reinforcement learning can be formally described by the Markov decision process framework below: consider the sequential decision model represented in Figure 2.1 [56], depicting the interaction between a decision-maker, or agent, and its environment. We assume that time is discrete, and that at each discrete time step $t \in \{0, 1, 2, \ldots, T\}$, the agent perceives the current state of the environment s_t from the set of all states S. We refer to T as the *horizon* and note that it may be either finite or infinite. On the basis of its state observation the agent selects an action a_t from the set of actions allowable in s_t, A_{s_t}. As a consequence, the following occurs immediately in the next time step: the agent receives a numerical signal r_{t+1} from the environment and the environment evolves to a new state s_{t+1} according to a probability distribution induced by s_t and a_t. The agent perceives state s_{t+1} and the interaction between agent and environment continues in this manner, either indefinitely or until some specified termination

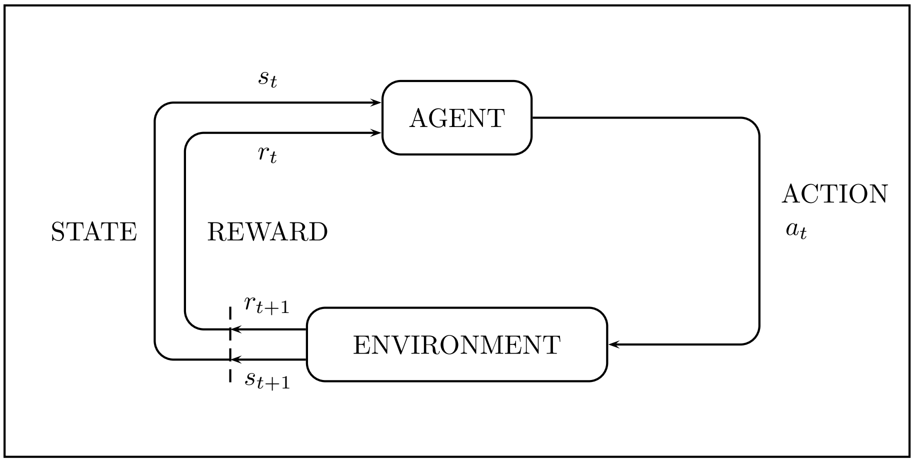

FIG. 2.1. *Agent-environment interaction*

point has been reached, in accordance with the length of the horizon. Here, we think of r_{t+1} as a means of providing the agent with a reward or a punishment as a direct consequence of its own actions, thereby enabling it to learn which action-selection strategies are good and which are bad via its own behaviour.

We further suppose that the following conditions are true of the stochastic nature of the environment: state transition probabilities obey the *Markov property*:

$$Pr(s_{t+1} = s|s_0, a_0, s_1, a_1, \ldots, s_t, a_t) = Pr(s_{t+1} = s|s_t, a_t)$$

and are *stationary*; that is, independent of time:

$$\text{for every } t \in T, Pr(s_{t+1} = s'|s_t = s, a_t = a) = P^a_{ss'}$$

The state and action spaces together with the transition probabilities and numerical rewards specified above comprise a discrete-time *Markov decision process*. Formally, we have the following:

DEFINITION 2.1. *A* finite Markov decision process *is a quadruple*

$$(S, \{A_s|s \in S\}, \{P(\cdot|s, a)|s \in S, a \in A_s\}, \{r(s, a)|s \in S, a \in A_s\})$$

where:

- *S is a finite set of states,*
- *$A = \cup_{s \in S} A_s$ is a finite set of actions,*
- *for every $s \in S, A_s$ is the set of actions allowable in state s,*
- *for every $s \in S$ and $a \in A_s, P(\cdot|s, a) : S \to [0, 1]$ is a stationary Markovian probability transition function; that is, for every $s' \in S$, $P(s'|s, a)$ is the probability of transitioning from state s to state s' under action a and will be denoted by $P^a_{ss'}$, and*
- *for every $s \in S$ and $a \in A_s$, $r(s, a)$ is the immediate reward associated with choosing action a in state s, and will be denoted by r^a_s.*

We frequently take $A_s = A$, that is, all actions are allowable in all states, and write a finite Markov decision process as (S, A, P, r).

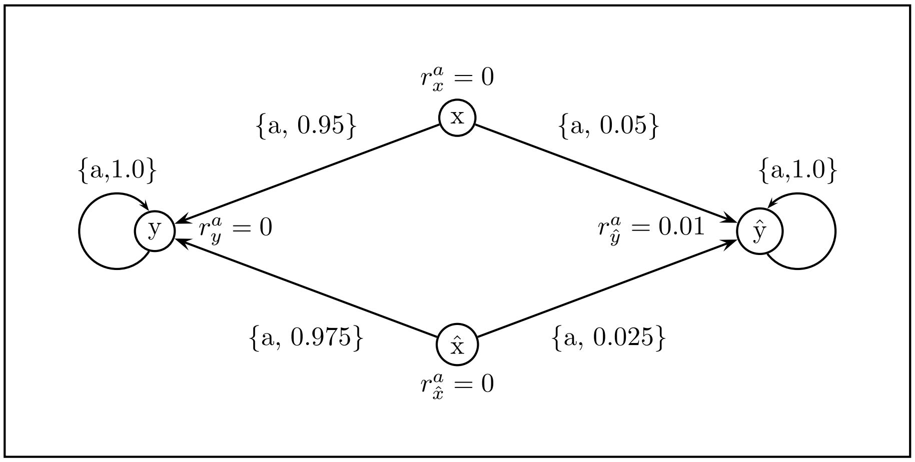

Fig. 2.2. *State transition diagram for a simple finite MDP*

A finite Markov decision process can also be specified via a state-transition diagram; Figure 2.2, for example, depicts a finite MDP with 4 states and 1 action.

A *Markov Decision Problem* consists of an MDP together with some optimality criterion concerning the strategies that an agent uses to pick actions. The particular Markov decision problem we will be concerned with is known as the *infinite-horizon expected discounted return reinforcement learning task.*

An action selection strategy, or *policy*, is essentially a mapping from states to actions, i.e. a policy dictates what action should be chosen for each state. More generally, one allows for policies that are stochastic, history-dependent, and even non-stationary. Here we will restrict our attention to randomized stationary Markov policies. Formally, a policy is a mapping $\pi : S \times A \to [0,1]$, such that $\pi(s,\cdot)$ is a probability distribution on A for each $s \in S$.

The optimality criterion of the Markov decision problems is concerned with finding a policy that maximizes the sum of the sequence of numerical rewards obtained through the agent's interaction with its environment. The most common optimality criterion, the infinite horizon total discounted reward task, involves finding a policy π that maximizes for every state $s \in S$, $\lim_{T\to\infty} \mathbb{E}^{\pi}[R_t | s_t = s]$ where $R_t = \sum_{k=0}^{T-(t+1)} \gamma^k r_{t+k+1}$ for some $\gamma \in [0,1)$ and $\mathbb{E}^{\pi}$ is the expectation taken with respect to the system dynamics following policy π. Such a maximizing policy is said to be *optimal*. Another optimality criterion is the average reward criterion, wherein one seeks to maximize for every state the cumulative sum of rewards averaged over the length of the horizon.

The total discounted reward criterion involves geometrically discounting the reward sequence. The intuition is that rewards obtained in the future are less valuable than rewards received immediately, an idea prevalent in economic theory; here the discount factor can be interpreted as a kind of interest rate. Another point of view comes from population modeling, where the discount factor γ can be viewed as the probability of an individual surviving to the next stage (and the process dies off with probability $1-\gamma$). Alternatively, we may simply view it as a mathematical tool to ensure convergence. In any case, the discounted reward model possesses many nice properties, such as a simplified mathematics in comparison to other proposed optimality criteria and existence of stationary optimal policies [53]. For this reason, it is the dominant criterion used for reinforcement learning tasks, and we concentrate on it in this work.

The expression $\lim_{T\to\infty} \mathbb{E}^\pi[R_t|s_t = s]$ that we seek to maximize in the infinite horizon discounted model is known as the *value* of a state s under a policy π, and is denoted $V^\pi(s)$. For finite MDPs, rewards are necessarily uniformly bounded; hence, the limit always exists and we may rewrite $V^\pi(s)$ as $\mathbb{E}^\pi[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}]$. The induced map on states, V^π, is called the *state-value function* (or simply *value function*) for π. Much research is concerned with estimating these value functions, as they contain key information towards determining an optimal policy.

In terms of value functions, a policy π^* is optimal if and only if $V^{\pi^*}(s) \geq V^\pi(s)$ for every $s \in S$ and policy π. As previously mentioned, an important fact about infinite horizon discounted models for finite MDPs is that an optimal policy always exists.

Given policy π, one can use the Markov property to derive for any $s \in S$,

$$V^\pi(s) = \sum_{a\in A_s} \pi(s,a)(r_s^a + \gamma \sum_{s'\in S} P^a_{ss'} V^\pi(s')) \tag{2.1}$$

The linear equations in 2.1 are known as the *Bellman equations* for policy π, and V^π is their unique solution. Note that while the value function for a given policy is unique, there may be many policies corresponding to the same value function.

The *optimal value function* V^*, corresponding to an optimal policy π^*, satisfies a more specialized family of fixed point equations,

$$V^*(s) = \max_{a\in A_s} (r_s^a + \gamma \sum_{s'\in S} P^a_{ss'} V^*(s')) \text{ for each } s \in S \tag{2.2}$$

of which it is the unique solution (see §6.1 and §6.2 of [53]). These are known as the *Bellman optimality equations.*

It is worth remarking that the existence and uniqueness of the solutions in these Bellman equations can be obtained from the Banach Fixed Point Theorem by applying the appropriate contraction mapping over the space of bounded real-valued functions on S equipped with the metric induced by the uniform norm (see Theorem 2.26 in § 2.4).

The Bellman equations are an important tool for reasoning about value functions and policies. They allow us to represent a value function as a limit of a sequence of iterates, which in turn can be used as the basis for dynamic programming algorithms for value function computation. Once more as a consequence of the Banach Fixed Point Theorem, one obtains:

THEOREM 2.2 (Policy Evaluation). *Given a randomized stationary policy π on a finite Markov decision process (S, A, P, r), define*

- $V_0^\pi(s) = 0$ *for every* $s \in S$ *and*
- $V_{i+1}^\pi(s) = \sum_{a\in A_s} \pi(s,a)(r_s^a + \gamma \sum_{s'\in S} P^a_{ss'} V_i^\pi(s'))$ *for every* $i \in \mathbb{N}$ *and* $s \in S$.

Then $(V_i^\pi)_{i\in\mathbb{N}}$ *converges to* V^π *uniformly.*

THEOREM 2.3 (Value Iteration). *Given a finite Markov decision process* (S, A, P, r). *Define*

- $V_0(s) = 0$ *for every* $s \in S$ *and*
- $V_{i+1}(s) = \max_{a\in A_s} (r_s^a + \gamma \sum_{s'\in S} P^a_{ss'} V_i(s'))$ *for every* $i \in \mathbb{N}$ *and* $s \in S$.

Then $(V_i)_{i\in\mathbb{N}}$ *converges to* V^* *uniformly.*

These results allow one to compute value functions up to any prescribed degree of accuracy. For example, if one is given a positive ϵ then iterating until the maximum difference between consecutive iterates is $\frac{\epsilon(1-\gamma)}{2\gamma}$ guarantees that the current iterate differs from the true value function by at most ϵ [53].

One can thus use value functions in order to compute optimal policies. For example, once one has performed value iteration, one can then determine an optimal policy by choosing for each state the action that maximizes its optimal value in the Bellman optimality equation, i.e.

$$\pi(s,a) \leftarrow \arg\max_{a\in A}(r_s^a + \gamma \sum_{s'\in S} P_{ss'}^a V^*(s')).$$

In practice, however, the optimal policy may stabilize for a given optimal value iterate long before the optimal value function itself has converged; in this case, the remaining iterations would serve only to waste time. As an alternative, one can instead iterate over policies. Given an arbitrary policy π, one can use policy evaluation to compute V^π and thereby obtain a measure of its quality. One can then attempt to improve π to π' by setting

$$\pi'(s,a) \leftarrow \arg\max_{a\in A}(r_s^a + \gamma \sum_{s'\in S} P_{ss'}^a V^\pi(s'));$$

this is known as policy improvement. If there is no improvement, that is, the policy is stable, then the policy is optimal; otherwise, one may continue to iterate in this manner. This is known as *policy iteration*: starting from an initial policy, one repeated performs policy evaluation and policy improvement until a stable optimal policy is achieved.

These dynamic programming algorithms constitute a standard MDP solution method; many alternative solution methods are based on them while aiming to improve computational efficiency. The problem with dynamic programming algorithms is that they are subject to the *curse of dimensionality*: a linear increase in state-space dimension leads to an exponential increase in running time. In general, such methods are impractical when dealing with large state spaces.

One typical method for overcoming such problems is state aggregation: one clusters together groups of states in some manner and defines a smaller MDP over the set of clusters. The hope is that one can recover a solution to the original MDP by solving the reduced model. However, clustering together states with different reward and probability parameters can be detrimental. We are thus led to the problem of how one should cluster states so as to recover good solutions; more generally, how does one best assess the quality of a state aggregation? The solution we propose is to use bisimulation metrics.

2.2. Discrete Bisimulation Metrics. Let $(S, A, \{P_{ss'}^a | s, s' \in S, a \in A\}, \{r_s^a | s \in S, a \in A\})$ be a given finite MDP. When should two states be placed in the same cluster of a state aggregation? Equivalently, what is the best state equivalence for MDP model reduction?

Givan, Dean and Greig [32] investigated several notions of MDP state equivalence for MDP model minimization: action-sequence equivalence, optimal value equivalence, and bisimulation. Two states are deemed action-sequence equivalent if for any fixed finite sequence of actions, their distributions over reward sequences are the same. Here let us remark that for any state, a fixed finite sequence of actions of length n induces a probability distribution over reward sequences of size n by means of the MDP's system dynamics. As [32] note, the problem with action-sequence equivalence is that it may equate states with different optimal values. To overcome such a limitation, the authors consider optimal value equivalence, wherein states are deemed equivalent if they have the same optimal value. Here again, however, problems arise: states deemed equivalent under optimal value equivalence may have markedly different MDP dynamics; in particular, they may have different optimal actions under an optimal policy and so are unsuitable for clustering. The authors go on to argue that bisimulation, a refinement of the first two equivalences, is the best state equivalence for model minimization.

Bisimulation has its origins in the theory of concurrent processes [50]. Milner [44] utilized strong bisimulation as a notion of process equivalence for his Calculus of Communicating Systems (CCS), a language used to reason about concurrent processes. Bisimulation in this context can informally be seen as a type of matching relation, i.e. processes p and q are related iff for every a-labeled transition that process p can make to process p', process q can make an a-labeled transition to some process q' related to p', and vice versa. A remarkable theorem shows that bisimulation equivalence on processes can be characterized by a modal logic known as *Hennessy-Milner logic* [36]; two processes are bisimilar if and only if they satisfy precisely the same formulas.

Remarkably, there was a precursor to the notion of bisimulation already available in the theory of Markov chains; this was called *lumpability* [38]. It did not use the fixed-point formulation and it did not make any connection with logic but, as its name suggests, it had the germ of the idea of probabilistic bisimulation well before bisimulation appeared in concurrency theory. Larsen and Skou [41] extended the notion of bisimulation to a probabilistic framework. Their *probabilistic bisimulation* was developed as an equivalence notion for labeled Markov chains (LMCs). They define probabilistic bisimulation both in terms of a maximal matching relation and establish a logical characterization result using a probabilistic modal logic. The definition of bisimulation by [32] is a simple extension of probabilistic bisimulation:

Definition 2.4. *Let (S, A, P, r) be a finite Markov decision process. A* stochastic bisimulation relation *R is an equivalence relation on S that satisfies the following property:*

$$sRs' \iff \textit{for each } a \in A, (r^a_s = r^a_{s'} \textit{ and for each } C \in S/R, P^a_s(C) = P^a_{s'}(C))$$

where $P^a_s(C) = \sum_{c \in C} P^a_{sc}$.

We say states s and s' are bisimilar*, written $s \sim s'$, iff sRs' for some stochastic bisimulation relation R.*

In other words, bisimilarity is the largest bisimulation relation on S, and roughly speaking, two states s and s' are bisimilar if and only if for every transition that s makes to a class of states, s' can make the same transition with the same probability and achieve the same immediate reward; and vice versa.

Bisimilarity was originally formulated by Park using fixed point theory [45]. This has been also done for probabilistic bisimilarity [58, 18] and for finite MDPs [24]. Note that the existence of a greatest fixed point in the definition below is guaranteed by an elementary theorem which asserts that a monotone function on a complete lattice has a greatest fixed point[1]:

Definition 2.5. *Let (S, A, P, r) be a finite Markov decision process, and let $\mathfrak{Rel}$ be the complete lattice of binary relations on S. Define $\mathcal{F} : \mathfrak{Rel} \to \mathfrak{Rel}$ by*

$$s\mathcal{F}(R)s' \iff \textit{for every } a \in A, (r^a_s = r^a_{s'} \textit{ and for each } C \in S/R_{rst}, P^a_s(C) = P^a_{s'}(C))$$

where R_{rst} is the reflexive, symmetric, transitive closure of R.
Then s and s' are bisimilar *iff $s \sim s'$ where $\sim$ is the greatest fixed point of $\mathcal{F}$.*

In the finite case, the operator $\mathcal{F}$ can be used to compute the bisimilarity partition: starting from an initial equivalence relation, the universal relation $S \times S$, iteratively apply $\mathcal{F}$ until a fixed point is reached. As each application of $\mathcal{F}$ either adds cluster-states or results in a fixed point, and there are only finitely many states, this procedure must stop.

[1] This is sometimes erroneously called the Knaester-Tarski Fixed Point Theorem. That is, however, a much more general theorem asserting that the fixed points of a monotone function on a complete lattice form a complete lattice.

Unfortunately, as an exact equivalence, bisimilarity suffers from issues of instability; that is, slight numerical differences in the MDP parameters, $\{r^a_s : s \in S, a \in A\}$ and $\{P^a_{ss'} : s, s' \in S, a \in A\}$, can lead to very different bisimilarity partitions. Consider the sample MDP in Figure 2.3 with 4 states labeled x, $\hat{x}$, y, and $\hat{y}$, and 1 action labeled a. Suppose $r^a_{\hat{y}} = 0$. Then all states share the same immediate reward and transition amongst themselves with probability one. So all states are bisimilar. On the other hand, if $r^a_{\hat{y}} > 0$ then $\hat{y}$ is the only state in its bisimulation class since it is the only one with a positive reward. Moreover, x and $\hat{x}$ are bisimilar if and only if they share the same probability of transitioning to $\hat{y}$'s bisimilarity class. Each is bisimilar to y if and only if that probability is zero. Thus, y, x, and $\hat{x}$ are not bisimilar to $\hat{y}$, $x \sim \hat{x}$ if and only if $p = p'$, $x \sim y$ if and only if $p = 1.0$, and $\hat{x} \sim y$ if and only if $p' = 1.0$. This example demonstrates that bisimilarity is simply too brittle; if $r_{\hat{y}}$ is just slightly positive, and p differs only slightly from p' then we should expect x and $\hat{x}$ to be practically bisimilar. However, an equivalence relation is too crude to capture this idea. To get around this, one generalizes the notion of bisimilarity equivalence through bisimulation metrics.

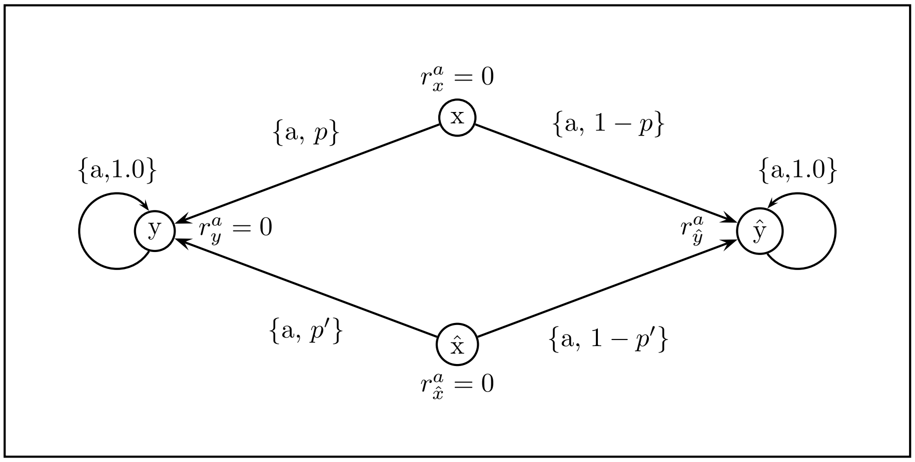

Fig. 2.3. *MDP demonstrating bisimilarity is too brittle*

Metrics can be used to give a quantitative notion of bisimulation that is sensitive to variations in the rewards and probabilistic transitions of an MDP. In [27, 28] we provided the following metric generalization of bisimulation for finite MDPs. Results appear here in slightly modified form:

Theorem 2.6. *Let (S, A, P, r) be a finite MDP and let $c \in (0, 1)$ be a discount factor. Let $\mathfrak{met}$ be the space of bounded pseudometrics on S equipped with the metric induced by the uniform norm. Define $F : \mathfrak{met} \to \mathfrak{met}$ by*

$$F(h)(s, s') = \max_{a \in A}((1 - c)|r^a_s - r^a_{s'}| + cT_K(h)(P^a_s, P^a_{s'}))$$

Then :

1. *F has a unique fixed point ρ^*,*
2. *$\rho^*(s, s') = 0 \iff s \sim s'$, and*
3. *for any $h_0 \in \mathfrak{met}$, $\|\rho^* - F^n(h_0)\| \leq \frac{c^n}{1-c}\|F(h_0) - h_0\|$.*

Here $T_K(h)(P,Q)$ is the Kantorovich probability metric[2]applied to finite distributions P and Q. We will introduce it in more generality in § 2.4.6 once we have set down some important concepts in continuous mathematics. For now, it is sufficient to note that in the finite case, it reduces to the following linear program:

$$\max_{u_i} \sum_{i=1}^{|S|} (P(s_i) - Q(s_i))u_i$$
$$\text{subject to: for every } i, j, u_i - u_j \leq h(s_i, s_j)$$

It can also be specified by the dual linear program

$$\begin{aligned} &\min_{\lambda_{kj}} \sum_{k,j=1}^{|S|} \lambda_{kj} h(s_k, s_j) \\ \text{subject to: } &\text{for every } k, \sum_j \lambda_{kj} = P(s_k) \\ &\text{for every } j, \sum_k \lambda_{kj} = Q(s_j) \\ &\text{for every } k, j, \lambda_{kj} \geq 0 \end{aligned}$$

which can be rewritten as $\min_\lambda \mathbb{E}_\lambda[h]$ where λ is a joint probability function on $S \times S$ with projections P and Q. This discrete minimization program has an interpretation as a *Hitchcock transportation problem*, an instance of the minimum-cost flow network optimization problem as seen in Figure 2.4.

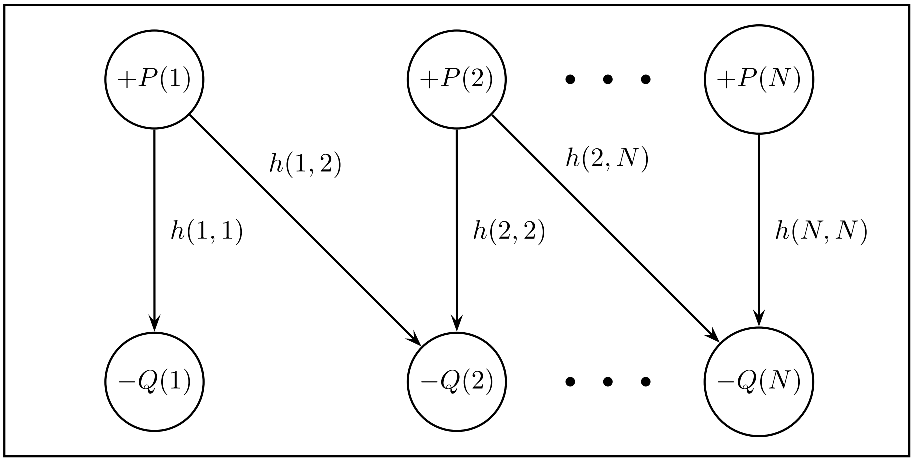

FIG. 2.4. *Hitchcock network transportation problem ($N = |S|$)*

Here we have $|S|$ source nodes and $|S|$ sink nodes. For each $s \in S$, there exists a source node labeled with a supply of $P(s)$ units and a sink node labeled with a demand (or negative supply) of

[2] Frustratingly, this metric likes to hide under a variety of names: Monge-Kantorovich, Kantorovich-Rubinstein, Hutchinson, Mallows, Wasserstein, Vasserstein, Earth Mover's Distance, Fortet-Mourier, and Dudley, to name a few.

$Q(s)$ units. Between each source node and each sink node, labelled respectively $P(s)$ and $Q(s')$ for some s, $s' \in S$, there is a transportation arc labelled with the cost of transporting one unit from the source to sink, given here by $h(s, s')$. A flow is an assignment of the number (nonnegative) of units to be shipped along all arcs. One requires that the total flow exiting a source node be equal to the supply of that node, and the total flow entering a sink node be equal to the demand at that node. One also requires that the total supply equals the total demand, which in this case is 1. The cost of a flow along an arc is simply the cost along that arc multiplied by the flow along that arc. The cost of the flow for the entire network is taken to be the sum of the flows along all arcs. The goal then is to find a flow of minimum cost.

There exist strongly polynomial algorithms to compute the minimum-cost flow problem [47, 61]. Therefore the Kantorovich metric in the discrete case can be computed in polynomial time, assuming of course that the pseudometric h is itself computable.

The key property of the Kantorovich metric is that it matches distributions, that is, assigns them distance zero only when they agree on the equivalence classes induced by the kernel of the underlying pseudometric cost function (see Lemma 3.7 in § 3). Therefore, it is not surprising that it can be used to capture the notion of bisimilarity, which requires that probabilistic transitions agree on bisimilarity equivalence classes.

Let us conclude with an example of the metric distances applied to the MDP in Figure 2.3. Using uniqueness of ρ^* and the identity $T_K(\rho^*)(\delta_x, \delta_y) = \rho^*(x, y)$ along with the fact that there is only one action, it is not hard to see that solving for ρ^* in the fixed point equations amounts to solving a set of linear equations. We therefore find:

$$\begin{aligned} \rho^*(x, \hat{x}) &= c|p - p'|r^a_{\hat{y}} & \qquad \rho^*(y, \hat{y}) &= r^a_{\hat{y}} \\ \rho^*(x, y) &= c(1 - p)r^a_{\hat{y}} & \qquad \rho^*(x, \hat{y}) &= (1 - cp)r^a_{\hat{y}} \\ \rho^*(\hat{x}, y) &= c(1 - p')r^a_{\hat{y}} & \qquad \rho^*(\hat{x}, \hat{y}) &= (1 - cp')r^a_{\hat{y}} \end{aligned}$$

Consider now the MDP in Figure 2.2. Even though states x and $\hat{y}$ are not bisimilar, we see that for any c they have ρ^*-distance $0.01 - 0.0095c$, which is much less than the maximum possible distance of 1; that is, they are very close to being bisimilar.

The most important property of the metrics is that they show that similar states have similar optimal values, and this relation varies smoothly with similarity. Formally, the optimal value function is continuous with respect to the state-similarity metrics.

Theorem 2.7 ([27]). *Let (S, A, P, r) be a finite MDP, $c \in (0, 1)$ be a metric discount factor, $\gamma \in [0, 1)$ be a reward discount factor, and ρ^* be the bisimulation metric given by Theorem 2.6. Suppose $\gamma \leq c$. Then V^* is $\frac{1}{1-c}$-Lipschitz continuous with respect to ρ^*, that is,*

$$|V^*(s) - V^*(s')| \leq \frac{1}{1 - c}\rho^*(s, s').$$

We can use this result to relate the optimal values of a state and its representation in an approximant by considering the original model and its approximant as one MDP.

2.3. Computing Bisimulation Metrics. We were able to compute the bisimulation metric by hand for the simple MDP pictured in Figure 2.3; but what can we say in the general case? In fact, the fixed point nature of the metrics permits the use of a dynamic programming algorithm in a manner analogous to the computation of the optimal value function: starting with the everywhere-zero metric, denoted by $\perp$, we iteratively apply the fixed point functional F until a desired level of accuracy is achieved. Since, as we noted, the Kantorovich operator can be computed in strongly polynomial time, we have an algorithm to calculate the state-similarity metrics - though one subject to the same shortcomings as traditional MDP dynamic programming algorithms. As only the distances are changing (and in fact converging) in the Kantorovich operator, and this object is itself an instance of a minimum-cost flow linear program, one immediately applicable speedup is to use cost re-optimization: that is, we can save the optimizing solutions for each Kantorovich linear program between iterations and use them to begin the Kantorovich linear program in the next iteration. The same idea was used in [64] to re-compute optimal network flows in the context of computing probabilistic simulations for probabilistic automata. As in that work, we are thereby saving on computation time at the cost of larger space requirements. This appears slightly more promising; but, can we do better? Indeed: a promising approach to quick and efficient approximation of the distances arises from the area of statistical sampling.

Suppose P and Q are approximated using the empirical distributions P_i and Q_i; that is, we sample i points $X_1, X_2, \dots, X_i$ independently according to P and define P_i by $P_i(x) = \frac{1}{i}\sum_{k=1}^{i} \delta_{X_k}(x)$. Similarly, write $Q_i(x) = \frac{1}{i}\sum_{k=1}^{i} \delta_{Y_k}(x)$. Remark that both P_i and Q_i are random variables defined over some ambient probability space. Then

$$T_K(h)(P_i, Q_i) = \min_{\sigma} \frac{1}{i} \sum_{k=1}^{i} h(X_k, Y_{\sigma(k)}) \tag{2.3}$$

where the minimum is taken over all permutations σ on i elements (see p. 5 of [60]). Now the Strong Law of Large Numbers tells us that both $(P_i(x))_1^\infty$ and $(Q_i(x))_1^\infty$ converge almost surely to $P(x)$ and $Q(x)$. Let us write $T_K^i(h)(P,Q)$ for $T_K(h)(P_i, Q_i)$ when the empirical distributions are fixed. Then as a consequence of the Strong Law of Large Numbers, $(T_K^i(h)(P,Q))_1^\infty$ converges to $T_K(h)(P,Q)$ almost surely; moreover replacing T_K by T_K^i in F yields a pseudometric,

$$\rho_i^*(s, s') = \max_{a \in A}((1-c)|r_s^a - r_{s'}^a| + cT_K^i(\rho_i^*)(P_s^a, P_{s'}^a)),$$

which converges almost surely to ρ^* as i gets large [26].

The importance of this result stems from the fact that the expression in equation (2.3) is an instance of the assignment problem from network optimization. This is a specialized network flow problem in which the underlying network is bipartite and all flow assignments are either 0 or 1. In graph-theoretic terminology, this is the problem of optimal matching in a weighted bipartite graph. Its specialized structure allows for fast, simple solution methods. For example, the Hungarian algorithm runs in worst case time $O(i^3)$, where i is the number of samples. Still, is the resulting sampling algorithm for estimating bisimulation distances really any better than the exact algorithms?

We have compared the Monte Carlo algorithm for a fixed number of samples along with the algorithms presented above, in terms of computational resources (space and time), and use in aggregation [26]. For purposes of illustration, we present here some of these results.

Experiments were run on MDPs given by an $n \times n$ grid world with two actions (move forward and rotate) and a single reward in the center of the room for $n = 3$, 5, and 7, and a flattened out version of the coffee robot MDP [8] in which a robot has to get coffee for a user while having to avoid getting wet. Each state in the grid world encodes both position as well as orientation of the agent; thus, the gridworld MDPs have 36, 100, and 196 states respectively. Additionally, the actions are deterministic. The coffee domain has 64 states and 4 actions, some with stochastic effects. For each domain, we computed: $\frac{1}{1-c}\rho^*$, the same with cost re-optimization, and $\frac{1}{1-c}\rho_i^*$ via sampling.

Exact computation of the Kantorovich metric in the first two methods was carried out using the MCFZIB minimum-cost flow solver [30]. An implementation of the Hungarian algorithm for the assignment problem was used to estimate the Kantorovich distances in the third method.

For each MDP, 10 transitions were sampled for each state and action, and this vector of samples was then used to estimate the empirical distribution throughout the whole run. The distance metric was obtained by averaging the distances obtained over 30 independent runs of this procedure.

Lastly, metrics were computed using three different values for the discount factor, here taking the metric and value discount factors to be the same, i.e. $c = \gamma$ with $\gamma \in \{0.1, 0.5, 0.9\}$.

Table 2.1 summarizes the running times in seconds for each method with the different discount factors. A ‘-’ means that the algorithm failed to compute the metric.

	Kantorovich	Re-optimized	Stochastic
3x3 gridWorld			
$\gamma = 0.1$	2.067	1.563	5.883
$\gamma = 0.5$	5.223	2.944	14.406
$\gamma = 0.9$	41.089	15.231	85.725
5x5 gridWorld			
$\gamma = 0.1$	-	-	44.200
$\gamma = 0.5$	-	-	109.473
$\gamma = 0.9$	-	-	653.645
7x7 gridWorld			
$\gamma = 0.1$	-	-	168.853
$\gamma = 0.5$	-	-	419.735
$\gamma = 0.9$	-	-	2625.16
Coffee Robot			
$\gamma = 0.1$	57.640	-	72.823
$\gamma = 0.5$	137.129	-	165.687
$\gamma = 0.9$	1024.42	-	1037.03

Table 2.1
Running times in seconds for different metric algorithms

We also compared the amount of space used by each method. This was measured using the *massif* tool of valgrind (a tool library in Linux). Table 2.2 presents the maximum number of bytes used by each algorithm when computing the distances for each MDP; an ‘*’ indicates an algorithm terminated prematurely due to maximum memory usage. In those cases where all algorithms were able to run to completion, the Monte Carlo algorithm either outperformed or performed comparably to the exact algorithms. Moreover, we compared the quality of the estimated distances with that of the exact distances by using each in simple aggregations schemes - and here too results were comparable [26]. All in all, when considering the tradeoff between the computational requirements of time and space, and the quality of the results, the Monte Carlo algorithm for calculating bisimulation distances significantly outperforms the others. Therefore, extending this sampling algorithm is the most promising approach to providing practical quantitative state-similarity for continuous Markov decision processes.

	Kantorovich	Re-optimized	Stochastic
3x3 gridWorld	80Mb	180Mb	80Kb
5x5 gridWorld	$1.8Gb^*$	$1.8Gb^*$	500Kb
7x7 gridWorld	$1.8Gb^*$	$1.8Gb^*$	1.8Mb
coffee robot	1.6Gb	$1.8Gb^*$	300Kb

Table 2.2
Memory usage in bytes for different metric algorithms

2.4. A Mathematical Review. Results will be stated without proof and can be found in most classical texts in probability and analysis, such as [55], [29], [20], and [4]. The subsections on metrics, convergence, topology, continuity and measure theory are elementary and can be skipped by a knowledgable reader; we include it just in case the reader wants to check our terminology. The subsection on probability metrics is perhaps less well known.

2.4.1. Metric Spaces. A metric is perhaps the simplest geometric structure that one can impose on a space. It is essentially a distance function; that is, a means of assigning a nonnegative numerical weight to pairs of points on a set in order to quantify how far apart they are.

Definition 2.8. *A* pseudometric *on a set S is a map $\rho : S \times S \to [0, \infty)$ such that for every s, s', s'' in S:*

1. $s = s' \Rightarrow \rho(s, s') = 0$
2. $\rho(s, s') = \rho(s', s)$
3. $\rho(s, s'') \leq \rho(s, s') + \rho(s', s'')$

If the converse of the first axiom holds as well, we say ρ is a metric*.*

A set S equipped with a metric (pseudometric) ρ is a metric (pseudometric) space*.*

Note that the kernel of a pseudometric when viewed as a real-valued function is an equivalence relation on S. We will denote the kernel of a pseudometric h on set S by $Rel(h)$.

Definition 2.9. *Given a pseudometric h on a set S, the equivalence relation $Rel(h)$ is defined by $sRel(h)s'$ if and only if $h(s, s') = 0$.*

A typical means of constructing a metric space is through a normed vector space, where one already has a notion of length of a vector through the norm function. Suppose $(V, \| \cdot \|)$ is such a space. Then $d(v, v') := \|v - v'\|$ is easily seen to define a metric on V.

A metric allows one to speak of the convergence of elements in a space: a sequence converges to a limit point if the distance between that limit point and the points in the sequence can eventually be made arbitrarily small.

Definition 2.10. *A sequence of elements $(x_n)_{n \in \mathbb{N}}$ in a metric space (S, ρ) converges to an element x in S if and only if for every positive ϵ there exists a natural number N, depending on ϵ, such that for every $n \geq N$, $\rho(x_n, x) < \epsilon$.*

As an example, whenever we speak of a sequence of bounded real-valued functions *converging uniformly*, we are implicitly invoking convergence in the space of bounded real-valued functions equipped with the metric induced by the uniform norm, i.e., $\|f\| := \sup_{x \in S} |f(x)|$.

Sometimes it is convenient to speak of the convergence of a sequence without having a definite candidate for its limit in mind. Suppose instead that we had considered a sequence whose pairwise distances could eventually be made arbitrarily small; we might expect that the sequence itself should converge. Unfortunately, such is not always the case.

Definition 2.11. *A sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space (S,ρ) is said to be* Cauchy *if and only if for every positive ϵ there exists a natural number N depending on ϵ such that for every $n, m \geq N$, $\rho(x_n, x_m) < \epsilon$.*

A metric space in which every Cauchy sequence converges is said to be *Cauchy-complete* or simply *complete*. An important example in this work consists of those pseudometrics on a set S that are bounded, i.e., any pseudometric h on S such that $\sup_{s,s'} |h(s,s')| < \infty$.

Completeness is just one of many special properties that can be attributed to a subset of a metric space. Here we consider a few more select sets and properties they might possess. First, given a point x in (S,ρ) and a fixed positive ϵ, we can consider all those points that are within ϵ-distance of x. These yield the *open and closed balls*, $B^\rho_\epsilon(x) = \{y \in S : \rho(x,y) < \epsilon\}$ and $C^\rho_\epsilon(x) = \{y \in S : \rho(x,y) \leq \epsilon\}$, respectively. More generally, a subset E of S is said to be *open* if for every point $e \in E$ there is some open ball $B^\rho_\epsilon(e)$ that is entirely contained in E. An open set containing x is also known as an *open neighborhood* of x. On the other hand, a subset F of S is said to be *closed* if its relative complement $S\backslash F$ is open. Closed subsets of a metric space can also be characterized by the following property: F is closed if and only if for every point x that is the limit of a convergent sequence in $F\backslash\{x\}$, x belongs to F, i.e. F contains all its limit points. Formally, a point p is a *limit point* of the set E if every open neighborhood of p contains some point of E other than p. This leads us to a type of subset useful for approximating the whole space. We say a subset X of S is *dense* in S if every point of S is a limit point of X or a point of X (or both). In particular, a metric space is said to be *separable* if it has some countable dense subset. In this work, we will be primarily interested in those metric spaces that are complete and separable, allowing us to work with an at most countably infinite set of points.

Definition 2.12. *A* Polish metric space *is a complete, separable metric space.*

From the point of view of approximating the whole space, there are two more interesting types of sets. A subset X is said to be *totally bounded* if for any positive ϵ it can be expressed as the union of finitely many open balls of radius ϵ. More generally, a subset X is *compact* if for every open cover of X, that is, for every collection of open subsets whose union contains X, there is a finite subcover of X. It is trivial to see that a totally bounded metric space is separable. More importantly, a metric space is compact if and only if it is totally bounded and complete. In particular, a compact metric space is Polish.

2.4.2. Topology.

This section is also elementary and can be skipped.

We note that different metrics can produce the same collection of open sets on a space, and that some properties depend only on this collection of open sets, rather than on a given metric. The set S equipped with a given collection of open sets is called a *topological space*.

Definition 2.13. *A collection $\mathcal{T}$ of subsets of a set S forms a* topology *on S if and only if:*

1. *The empty set $\emptyset$ and the whole set S belong to $\mathcal{T}$,*
2. *$\mathcal{T}$ is closed under finite intersections, i.e. if $\{U_i\}_{i=1}^n$ is a finite collection in $\mathcal{T}$ then $\bigcap_{i=1}^n U_i \in \mathcal{T}$, and*
3. *$\mathcal{T}$ is closed under arbitrary unions, i.e. if $\{U_\alpha\}_{\alpha\in J}$ is a collection in $\mathcal{T}$ for some index set J then $\bigcup_{\alpha\in J} U_\alpha \in \mathcal{T}$.*

A set S with a topology $\mathcal{T}$ is known as a topological space.

If $(S,\mathcal{T})$ is a topological space then a subset U of S is an *open set* of S if U belongs to $\mathcal{T}$ and a subset V of S is a *closed set* of S if its relative complement $X - V$ is open in S. Properties that refer only to the collection of open sets will be referred to as *topological.* It is not hard to show that for any metric space, the collection of open sets as defined in § 2.4.1 forms a topology called the *metric topology.*

Given two sets X and Y, we can form the cartesian product $X \times Y$. Naturally, if X and Y have associated topologies, we would like to associate a topology to $X \times Y$. The standard method for doing so uses the coordinate or projection maps on the product.

Definition 2.14. *Given the cartesian product $X \times Y$ of two sets X and Y, let $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ be defined by $\pi_1(x, y) = x$ and $\pi_2(x, y) = y$. The maps π_1 and π_2 are called the* projections *of $X \times Y$ onto its first and second coordinates, respectively.*

Definition 2.15. *A* subbasis *$\mathcal{S}$ for a topology on a set X is a collection of subsets of X whose union equals X. The* topology generated by the subbasis *$\mathcal{S}$ is the collection $\mathcal{T}$ of all unions of finite intersections of elements of $\mathcal{S}$.*

Definition 2.16. *Let X and Y be topological spaces. The product topology on $X \times Y$ is the topology generated by the subbasis $\mathcal{S} = \{\pi_1^{-1}(U) | U$ is open in $X\} \cup \{\pi_2^{-1}(V) | V$ is open in $Y\}$.*

In particular, if X and Y are metric spaces with metrics ρ_X and ρ_Y, respectively, then the *product metric* $\rho_{X \times Y}$ defined by $\rho_{X \times Y}((x_1, y_1), (x_2, y_2)) = \max\{\rho_X(x_1, x_2), \rho_Y(y_1, y_2)\}$ generates the product topology on $X \times Y$.

2.4.3. Continuity.

Continuity is a crucial property for our work on approximating spaces and functions on those spaces. Loosely speaking, a function is continuous if the output of the function cannot change too abruptly with small changes in its input.

Continuity in topological spaces is defined as follows:

Definition 2.17. *A function $f : (X, \mathcal{T}_X) \to (Y, \mathcal{T}_Y)$ be topological spaces is continuous if for each open set $O_Y \in \mathcal{T}_Y$, the preimage $f^{-1}(O_Y) \in \mathcal{T}_X$.*

Continuity is important for defining equivalence of topological spaces; two topological spaces are equivalent, or *homeomorphic*, if there exists a continuous bijection between them such that its inverse is also continuous.

Definition 2.18. *A* Polish space *is a topological space that is homeomorphic to a Polish metric space.*

Some important results can be established under weaker continuity conditions. One such condition is lower semicontinuity.

Definition 2.19. *Let $(X, \mathcal{T})$ be a topological space and let $f : X \to \mathbb{R} \cup \{-\infty, \infty\}$. Then f is* lower semicontinuous *if for each half-open interval of the form (r, ∞), the preimage $f^{-1}(r, \infty) \in \mathcal{T}$.*

Continuity in metric spaces is defined as follows:

Definition 2.20. *A function $f : (X, \rho_X) \to (Y, \rho_Y)$ between metric spaces is* continuous at a point *$x \in X$ if for every $\epsilon > 0$ there is a $\delta > 0$, depending on x and ϵ, such that for every $x' \in X$ with $\rho_X(x, x') < \delta$ we have $\rho_Y(f(x), f(x')) < \epsilon$.*

We say f is continuous *if it is continuous at every point of X.*

If the topologies $\mathcal{T}_X$ and $\mathcal{T}_Y$ are generated by metrics ρ_X and ρ_Y respectively, then defintion 2.17 and definition 2.20 coincide.

If the δ in definition 2.20 can be chosen so as to depend on ϵ alone, i.e. independent of the point x, then f is said to be *uniformly continuous.* A stronger form of uniform continuity is Lipschitz continuity, which plays an important part in this work.

Definition 2.21. *A function $f(X, \rho_X) \to (Y, \rho_Y)$between metric spaces is* Lipschitz continuous *if for some constant α, $\rho_Y(f(x), f(x')) \leq \alpha \rho_X(x, x')$ for every $x, x' \in X$.*

Any such constant α is known as a *Lipschitz constant* for this mapping; the greatest lower bound of all such Lipschitz constants is itself a Lipschitz constant, known as *the Lipschitz constant.* For either case, we will sometimes write that f is α-Lipschitz continuous.

Obviously every Lipschitz continuous function is uniformly continuous, and every uniformly continuous function is continuous, but the converse is not generally true in either case. For compact metric spaces, however, the situation is much more well-behaved. Here, every continuous function is indeed uniformly continuous. Moreover, if f is real-valued then it has a minimum value and a maximum value, each of which is attained.

Continuity in metric spaces can alternatively be characterized in terms of convergent sequences: f is continuous if for every convergent sequence $(x_n)_{n\in\mathbb{N}}$ in X with limit x, the sequence $(f(x_n))_{n\in\mathbb{N}}$ is convergent with limit $f(x)$. One can analogously defined a sequential version of lower semicontinuity.

Definition 2.22. *A function $f : (X, \rho) \to \mathbb{R} \cup \{-\infty, \infty\}$ on a metric space is* sequentially lower semicontinuous *if for any sequence $(x_n)_{n\in\mathbb{N}}$ converging to x in X, $\liminf_{n\to\infty} f(x_n) \geq f(x)$.*

Again, if the topology $\mathcal{T}$ on a space X is generated by the metric ρ then Definition 2.19 and 2.22 coincide. We will make more use of the sequential definitions of continuity.

One can analogously define f to be *upper semicontinuous* by requiring $\limsup_{n\to\infty} f(x_n) \leq f(x)$. It is easily seen that a real-valued function is continuous if and only if it is both lower semicontinuous and upper semicontinuous. The intuition behind these definitions is that semicontinuous functions allow for abrupt (discontinuous) jumps in one vertical direction; this can be seen through the prototypical examples of semicontinuous functions: the indicator function of an open set is always lower semicontinuous while the indicator function of a closed set is always upper semicontinuous. In this work, we will be particularly interested in lower semicontinuous functions due to several important properties; for example, the pointwise supremum of an arbitrary collection of uniformly bounded lower semicontinuous functions on a metric space is itself lower semicontinuous, and a lower semicontinuous function on a compact space attains its minimum. The statement of the following theorem can be found as part a) of Theorem B.5 in [53] and by noting that f is lower semicontinuous if and only if $-f$ is upper semicontinuous.

Theorem 2.23. *Let X be a Polish metric space, Y a compact subset of a Polish metric space, and $f : X \times Y \to \mathbb{R}$ be a lower semicontinuous function. Then $g : X \to \mathbb{R}$ defined by $g(x) = \min_{y\in Y} f(x, y)$ is lower semicontinuous on X.*

Sometimes we need to speak of continuity of a family of functions such that they collectively have equal variation over a given neighborhood.

Definition 2.24. *A family of functions $\mathcal{F}$ between metric spaces (X, ρ_X) and (Y, ρ_Y) is* equicontinuous at a point $x \in X$ *if for every $\epsilon > 0$ there is a $\delta > 0$, depending on x and ϵ, such that for every $x' \in X$ with $\rho_X(x, x') < \delta$ and for every $f \in \mathcal{F}$, we have $\rho_Y(f(x), f(x')) < \epsilon$.*

2.4.4. Fixed Point Theory.

Fixed point theory plays a major role in this paper. Here we recall some basic definitions and a theorem from fixed point theory on lattices, which can be found in any basic text [63].

Let $(L, \preceq)$ be a partial order. If it has least upper bounds and greatest lower bounds of arbitrary subsets of elements, then it is said to be a *complete lattice*. A function $f : L \to L$ is said to be *monotone* if $x \preceq x'$ implies $f(x) \preceq f(x')$. A point x in L is said to be a *prefixed point* if $f(x) \preceq x$, a *postfixed point* if $x \preceq f(x)$ and a *fixed point* if $x = f(x)$. The importance of these definitions arises in the following theorem.

Theorem 2.25 (Knaster-Tarski Fixed Point Theorem). *Let L be a complete lattice, and suppose $f : L \to L$ is monotone. Then f has a least fixed point, which is also its least prefixed point, and f has a greatest fixed point, which is also its greatest postfixed point.*

This is an elementary theorem sometimes called the Knaster-Tarski theorem in the literature. In fact the Knaster-Tarski theorem is a much stronger statement to the effect that the collection of fixed points is itself a complete lattice.

A more common fixed point theorem comes from the theory of metric spaces and has the advantage of being constructive in nature; its proof can be found in most basic texts in analysis, e.g. [55].

THEOREM 2.26 (Banach Fixed Point Theorem). *Suppose (X, d) is a complete metric space and $T : X \to X$ is a contraction mapping; that is, for some $c \in [0, 1)$*

$$d(Tx, Tx') \leq c \cdot d(x, x')$$

for every x, x' in X. Then:

1. *T has a unique fixed point, x^*, and*
2. *for any $x_0 \in X$, $d(x^*, T^n x_0) \leq \frac{c^n}{1-c} d(Tx_0, x_0)$.*

In particular, $\lim_{n\to\infty} T^n x_0 = x^*$.

2.4.5. Probability and Measure. A rather unfortunate consequence of moving to uncountably infinite state spaces is that we can no longer specify transition probabilities point-to-point; one needs to specify probabilities on sets of points and even then not all sets can be "measured" in this way.

DEFINITION 2.27. *A* σ-algebra *or* σ-field *on a set S is a collection Σ of subsets of S satisfying the following axioms:*

1. *The empty set $\emptyset$ and the whole set S belong to Σ,*
2. *Σ is closed under complements, i.e. if $E \in \Sigma$ then $S \backslash E \in \Sigma$, and*
3. *Σ is closed under countable unions, i.e. if $(E_i)_{i=1}^{\infty}$ is a sequence in Σ then $\bigcup_1^{\infty} E_i \in \Sigma$.*

The members of Σ are known as the *measurable sets*. The pair (S, Σ) is known as a *measurable space*. Given a topological space $(S, \mathcal{T})$, there is a unique smallest σ-algebra $\mathcal{B}(\mathcal{T})$ that contains all the open sets; this is known as the *Borel σ-algebra*. Its members are said to be *Borel measurable* sets.

More generally, if $\mathcal{E}$ is any collection of subsets of a set S, there is a unique σ-algebra $\mathcal{M}(\mathcal{E})$ containing $\mathcal{E}$. It is called the *σ-algebra generated by $\mathcal{E}$*.

Given two spaces X and Y with associated σ-algebras, we can again form the cartesian product $X \times Y$ and associate to it a σ-algebra.

DEFINITION 2.28. *Let (X, Σ_X) and (Y, Σ_Y) be measurable spaces and let π_1 and π_2 be the coordinate maps defined in Definition 2.14. The* product σ-algebra $\Sigma_X \otimes \Sigma_Y$ *on $X \times Y$ is the σ-algebra generated by the set $\mathcal{E} = \{\pi_1^{-1}(E) | E \in \Sigma_X\} \cup \{\pi_2^{-1}(F) | F \in \Sigma_Y\}$.*

Now suppose that X and Y are two topological spaces. There are two ways of defining a σ-algebra on $X \times Y$: the Borel σ-algebra generated by the product topology, and the product σ-algebra on X and Y each equipped with its Borel σ-algebra. In general, these need not be equal. However, in the case of separable metric spaces, they are.

PROPOSITION 2.29. *Let X and Y be metric spaces and let $X \times Y$ be equipped with the product metric. If X and Y are separable then the product σ-algebra on $X \times Y$ is equal to the Borel σ-algebra of $X \times Y$.*

DEFINITION 2.30. *Given a measurable space (S, Σ), a* measure *is a set function $\mu : \Sigma \to [0, \infty]$ such that*

1. *$\mu(\emptyset) = 0$, and*
2. *for any pairwise disjoint sequence of sets $(E_i)_{i=1}^{\infty}$ in Σ, $\mu(\bigcup_1^{\infty} E_i) = \Sigma_1^{\infty} \mu(E_i)$.*

If μ take values in $[0, 1]$ then it is a subprobability measure*; if in addition $\mu(S) = 1$ then it is a* probability measure . *The triple (S, Σ, μ) is known as a* measure space *(respectively,* subprobability space*,* probability space*).*

Sometimes we need to assign weights of a probabilistic type to all subsets of a space, at the cost of losing some of the nice properties of a probability measure; such is frequently the case in the theory of empirical processes, where one cannot guarantee that all the sets one may encounter in practice will be measurable.

Definition 2.31. *An* outer probability measure *on a set S is a set function $\phi : 2^S \to [0,1]$ satisfying*

1. $\phi(\emptyset) = 0$*,*
2. $E \subset F$ *implies* $\phi(E) \leq \phi(F)$*, and*
3. *for any sequence* $(E_i)_{i=1}^{\infty}$ *of subsets of* S*,* $\phi(\bigcup_1^\infty E_i) \leq \Sigma_1^\infty \phi(E_i)$*.*

Every probability measure can be extended to an outer probability measure, and conversely, every outer probability measure can be used to construct a σ-algebra on which it is a probability measure. Note as well that any set of outer probability zero has complement with outer probability one.

Definition 2.32. *A probability measure on a metric space is* tight*, or* inner regular*, if it can be approximated from within by compact sets, that is, μ is tight if for every Borel measurable set E, $\mu(E) = \sup_K \mu(K)$ where the supremum is taken over all compact subsets K contained in E.*

Theorem 2.33 (Ulam's Tightness Theorem). *Every probability measure on a Polish metric space is tight.*

Measures can be extended to act on functions through the process of integration. We will assume the reader is familiar with the basic ideas of integration, if not the details, as the details are involved and add nothing to the exposition here. Suffice it to say that, just as only certain subsets can be measured, so too can only certain functions be integrated. Formally, a function f between measurable spaces (X, Σ_X) and (Y, Σ_Y) is said to be *measurable* if the preimage of every Σ_Y-measurable set is Σ_X-measurable, i.e. $\{f^{-1}(E) : E \in \Sigma_Y\} \subseteq \Sigma_X$. A real-valued function f on a measurable space (S, Σ) is measurable, or in the language of probability theory, a *random variable*, if it is measurable as just defined, where $\mathbb{R}$ is equipped with its usual Borel σ-field. The prototypical measurable functions are the *simple functions*: finite linear combinations of indicator functions on measurable sets. Real-valued measurable functions can be approximated in a nice way by simple functions.

Theorem 2.34 (p.47 of [29]). *Let (S, Σ) be a measurable space. If $f : S \to [0, \infty]$ is measurable, then there is a sequence $(\phi_n)_{n\in\mathbb{N}}$ of simple functions such that $0 \leq \phi_1 \leq \phi_2 \leq \cdots \leq f$, $(\phi_n)_{n\in\mathbb{N}}$ converges to f pointwise, and $(\phi_n)_{n\in\mathbb{N}}$ converges to f uniformly on any set on which f is bounded.*

If S is a metric space and Σ its Borel σ-field, then every continuous function on S is measurable. Given a sequence of measurable functions, its pointwise supremum, infimum, and limit (when it exists) are all measurable. Lastly, if the integral of the absolute value of a measurable function f with respect to a measure μ exists and is finite, then f is said to be *integrable.* The collection of all such f for a given μ is denoted by $L^1(\mu)$ (here it is standard to identify functions that differ on a set of μ-measure zero).

Let us now consider convergence of probability measures on a metric space. Since probability measures are essentially just set functions, it is natural to attempt to analyze their convergence properties through pointwise converge, that is, to say that a sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ converges to probability measure μ if $(\mu_n(E))_{n\in\mathbb{N}}$ converges to $\mu(E)$ for every measurable set E. However, such convergence is too strong: consider the Dirac measure δ_x, which assigns a value of 1 if and only if a given measurable set contains the point x and 0 otherwise. Take $[0,1]$ with its Borel σ-algebra and consider the sequence of Dirac measures on $\{\frac{1}{n} : n \in \mathbb{N}\}$. It would be quite natural to expect, if not demand, that this sequence converges to the Dirac measure at zero.

However, taking the Borel measurable singleton $\{0\}$ in the definition of pointwise convergence would yield $\lim_{n\to\infty} \delta_{\frac{1}{n}}(\{0\}) = 0 = \delta_0(\{0\}) = 1$, which is clearly not the case. It is not hard to show here that pointwise convergence over the measurable sets is equivalent to pointwise convergence over bounded measurable functions, that is, convergence of $(\mu_n(f))_{n\in\mathbb{N}}$ to $\mu(f)$ for every bounded measurable function f. Therefore, one way of weakening convergence is to consider a similar pointwise convergence, but over a smaller class of functions. Formally, we say that $\{\mu_n\}$ *converges weakly* to μ if $(\mu_n(f))_{n\in\mathbb{N}}$ converges to $\mu(f)$ for every bounded *continuous* real-valued function f. It is clear that the Dirac measures on $\{\frac{1}{n} : n \in \mathbb{N}\}$ do indeed converge weakly to the Dirac measure at 0.

THEOREM 2.35 ([51]). *Let X be a separable metric space and $(\mu_n)_{n\in\mathbb{N}}$ be any sequence of measures on X. Let $\mathcal{A}_0 \subseteq C(X)$ be a family of functions which is equicontinuous at every point $x \in X$ and uniformly bounded, that is, for some constant M, $|f(x)| \leq M$ for every $x \in X$ and $f \in \mathcal{A}_0$. Then $\mu_n \Rightarrow \mu$ if and only if*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{A}_0} \left| \int f d\mu_n - \int f d\mu \right| = 0.$$

2.4.6. Probability Metrics. There are numerous ways of defining a notion of distance between probability measures on a given space [31]. Two typical ones are the total variation distance, capturing strong convergence of probability measures, and the Kullback-Leibler divergence, capturing certain information-theoretic properties of the measures. Note that the Kullback-Leibler divergence fails to satisfy the symmetry and triangle inequality axioms for a metric. As previously mentioned, however, the particular probability metric of which we make use is known as the Kantorovich metric. Its use in defining metrics for bisimulation was first demonstrated by van Breugel and Worrell [58]. We present it here in greater generality; all results are taken from the books by Rachev and Rueschendorf [54] and Villani [60], unless otherwise stated.

DEFINITION 2.36. *Let S be a Polish metric space, h a bounded pseudometric on S that is lower semicontinuous on $S \times S$ with respect to the product topology, and $Lip(h)$ be the set of all bounded functions $f : S \to \mathbb{R}$ that are measurable with respect to the Borel σ-algebra on S and that satisfy the Lipschitz condition $f(x) - f(y) \leq h(x, y)$ for every $x, y \in S$. Let P and Q be probability measures on S. Then the* Kantorovich distance *$T_K(h)$ is defined by*

$$T_K(h)(P, Q) = \sup_{f\in Lip(h)} (P(f) - Q(f)).$$

The Kantorovich metric arose in the study of optimal mass transportation. The following description is due to Villani [60]: assume we are given a pile of sand and a hole, occupying measurable spaces (X, Σ_X) and (Y, Σ_Y), each representing a copy of (S, Σ) (figure 2.5). The pile of sand and the hole obviously have the same volume, and the mass of the pile is assumed to be normalized to 1. Let P and Q be measures on X and Y respectively, such that whenever $A \in \Sigma_X$ and $B \in \Sigma_Y$, $P[A]$ measures how much sand occupies A and $Q[B]$ measures how much sand can be piled into B. Suppose further that we have some measurable cost function $h : X \times Y \to \mathbb{R}$, where $h(x, y)$ tells us how much it costs to transfer one unit of mass from a point $x \in X$ to a point $y \in Y$. Here we consider h satisfying the conditions of Definition 2.36. The goal is to determine a plan for transferring all the mass from X to Y while keeping the cost at a minimum. Such a transfer plan is modelled by a probability measure λ on $(X \times Y, \Sigma_X \otimes \Sigma_Y)$, where $d\lambda(x, y)$ measures how

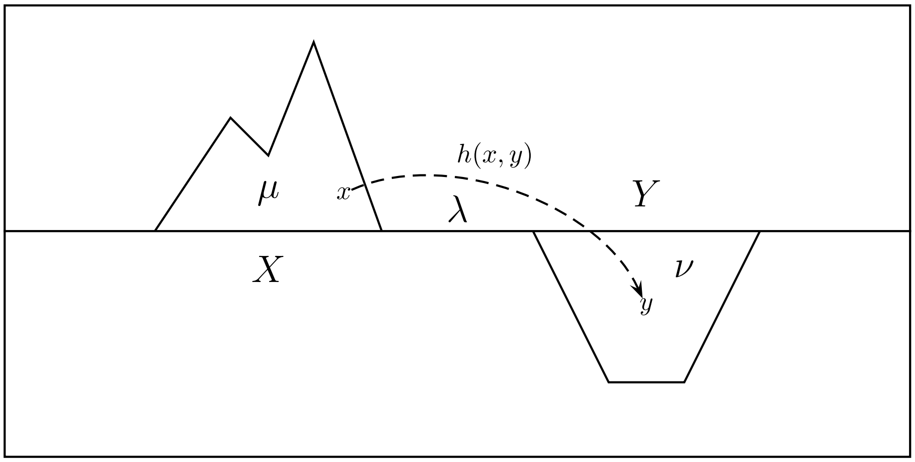

Fig. 2.5. *Kantorovich optimal mass transportation problem*

much mass is transferred from location x to y. Of course, for the plan to be valid we require that $\lambda[A \times Y] = P[A]$ and $\lambda[X \times B] = Q[B]$ for every measurable A and B. A plan satisfying this condition is said to have marginals P and Q, and we denote the collection of all such plans by $\Lambda(P, Q)$.

Definition 2.37. *Let S, P, and Q be as in Definition 2.36. Then $\Lambda(P, Q)$ consists of all measures on the product space $S \times S$ with marginals P and Q.* We can now restate the goal formally as:

$$\text{minimize } h(\lambda) \text{ over } \lambda \in \Lambda(P, Q)$$

This is actually an instance of an infinite linear program. Fortunately, under very general circumstances, it has a solution and admits a dual formulation.

Let us first note that measures in $\Lambda(P, Q)$ can equivalently be characterized as those λ satisfying:

$$P(\phi) + Q(\psi) = \lambda(\phi + \psi)$$

for every $(\phi, \psi) \in L^1(P) \times L^1(Q)$, where $\phi + \psi$ refers to the map that takes (x, y) to $\phi(x) + \psi(y)$. As a consequence of this characterization we have the following inequality:

$$\sup_{f \in Lip(h, C_b(S))} (P(f) - Q(f)) \leq T_K(h)(P, Q) \leq \inf_{\lambda \in \Lambda(P,Q)} h(\lambda) \tag{2.4}$$

where

Definition 2.38. *A function $f : S \to [0, \|h\|]$ on a topological space S belongs to $Lip(h, C_b(S))$ if and only if it is continuous and bounded on S (in fact, bounded by $\|h\|$) and 1-Lipschitz continuous with respect to h.*

Note that h need not generate the topology on S, and so Lipschitz continuity with respect to h does not immediately imply continuity on S.

The leftmost and rightmost terms in inequality (2.4) are examples of infinite linear programs in duality. It is a highly nontrivial result that there is no duality gap in this case (see for example Theorem 1.3 and the proof of Theorem 1.14 in [60]).

Theorem 2.39 (Kantorovich-Rubinstein Duality Theorem). *Assume the conditions of Definition 2.36, Definition 2.37, and Definition 2.38. Then there is no duality gap in equation 2.4, that is,*

$$T_K(h)(P,Q) = \sup_{f \in Lip(h, C_b(S))} (P(f) - Q(f)) = \inf_{\lambda \in \Lambda(P,Q)} h(\lambda) \tag{2.5}$$

Note that for any point masses δ_x, δ_y, we have $T_K(h)(\delta_x, \delta_y) = h(x,y)$ since $\delta_{(x,y)}$ is the only measure with marginals δ_x and δ_y. As a result, we obtain that any bounded lower semicontinuous pseudometric h can be expressed as $h(x,y) = \sup_{f \in \mathcal{F}} (f(x) - f(y))$ for some family of continuous functions $\mathcal{F}$ (we used this property at the end of § 2.2 to compute the state-similarity metric by hand for a very simple finite MDP).

Suppose P and Q are finite sums of Dirac measures assigning equal mass to each of n points, respectively, i.e. $P = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k}$ and $Q = \frac{1}{n}\sum_{k=1}^{n} \delta_{Y_k}$ for points $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ in S. Then the Kantorovich metric simplifies according to

$$T_K(h)(P,Q) = \min_{\sigma} \frac{1}{n} \sum_{k=1}^{n} h(X_k, Y_{\sigma(k)})$$

where the minimum is taken over all permutations σ on n elements. This is particularly useful for measuring the distance between empirical measures.

The Kantorovich metric also admits a characterization in terms of the coupling of random variables. We may write $T_K(h)(P,Q) = \min_{(X,Y)} \mathbb{E}[h(X,Y)]$ where the expectation is taken with respect to the joint distribution of (X,Y) and the minimum is taken with respect to all pairs of random variables (X,Y) such that the marginal distribution of X is P and the marginal distribution of Y is Q.

3. Bisimulation Metrics for Continuous MDPs. The first thing we have to deal with in moving to infinite state spaces[3] is the issue of measurability; simply put, we can no longer specify probabilities point-to-point. One needs to look at the probabilities of sets of states, and even then, not all sets can be measured in this way. Formally, we have a potentially uncountably infinite state space, S, equipped with a σ-algebra of measurable sets, Σ. We may think of Σ as providing some sort of "information resolution" - that is, the only pertinent sets of states are those that are measurable (and we ignore the rest). Following along these lines, we need to ensure that the reward and probability functions satisfy certain measurability conditions, that is, that they behave well with respect to measurable sets. Formally, we have the following:

Definition 3.1. *A Markov decision process (MDP) is a tuple (S, Σ, A, P, r), where (S, Σ) is a measurable space, A is a finite set of actions, $r : S \times A \to \mathbb{R}$ is a measurable reward function, and $P : S \times A \times \Sigma \to [0,1]$ is a labeled stochastic transition kernel, i.e.*

- *for every $a \in A$ and $s \in S$, $P(s,a,\cdot) : \Sigma \to [0,1]$ is a probability measure, and*
- *for every $a \in A$ and $X \in \Sigma$, $P(\cdot,a,X) : S \to [0,1]$ is a measurable function.*

We will use the following notation: for $a \in A$ and $s \in S$, P_s^a denotes $P(s,a,\cdot)$ and r_s^a denotes $r(s,a)$. Given measure P and integrable function f, we denote the integral of f with respect to P by $P(f)$.

We also make the following assumptions:

1. S is Polish space equipped with its Borel σ-algebra, Σ,

[3] We will still assume finitely many actions; what to do when this is not the case is beyond the scope of this work.

2. the image of r is contained in $[0,1]$

3. for each $a \in A$, $r(\cdot, a)$ is continuous on S.

4. for each $a \in A$, P^a_s is (weakly) continuous as a function of s, that is, if s_n tends to s in S then for every bounded continuous function $f : S \to \mathbb{R}$, $P^a_{s_n}(f)$ tends to $P^a_s(f)$.

Our presentation of bisimilarity here amounts to little more than a mild extension through the addition of rewards to the definition of bisimilarity given by [17] in their work on labelled Markov processes (LMPs).

Let R be an equivalence relation on S. We now have two notions of “visibility” on S: the measurable sets, as determined by the σ-algebra on S, and the sets built up from the equivalence classes of R. Naturally, we are interested in those sets that are visible under both criteria - measurability and equivalence. Let us formalize these concepts.

Definition 3.2. *Given a relation R on a set S, a subset X of S is said to be R-*closed *if and only if the collection of all those elements of S that are reachable by R from X, $R(X) = \{s' \in S | \exists s \in X,\ sRs'\}$, is itself contained in X.*

Definition 3.3. *Given a relation R on a measurable space (S, Σ), we write $\Sigma(R)$ for the set of those Σ-measurable sets that are also R-closed, $\{X \in \Sigma | R(X) \subseteq X\}$.*

When R is an equivalence relation then to say that a set X is R-closed is equivalent to saying that X is a union of R-equivalence classes. In this case $\Sigma(R)$ consists of those measurables that can be partitioned into R-equivalence classes.

Definition 3.4. *Let (S, Σ, A, P, r) be an MDP satisfying the conditions of Definition 3.1. An equivalence relation R on S is a* bisimulation relation *if and only if it satisfies*

$$sRs' \Leftrightarrow \text{ for every } a \in A,\ r^a_s = r^a_{s'} \text{ and for every } X \in \Sigma(R),\ P^a_s(X) = P^a_{s'}(X).$$

Bisimilarity *is the largest of the bisimulation relations.*

Note that it is not immediately clear that bisimilarity itself is a bisimulation relation (transitivity is not obvious); that this is indeed the case will be shown in the proof of Theorem 3.10 through a fixed point characterization of bisimilarity. By contrast, [17] prove transitivity through a logical characterization of bisimilarity.

As in Theorem 2.6, we will develop a metric anaologue of bisimilarity over a certain space of pseudometrics on S; here, however, continuity and measurability conditions come into play.

Definition 3.5. *Let S be a Polish space. Then we define $\mathfrak{met}$ to be the set of bounded pseudometrics on a S equipped with the metric induced by the uniform norm. We define $\mathfrak{lsc}_\mathfrak{m}$ to be the set of bounded pseudometrics on S that are lower semicontinuous on $S \times S$ endowed with the product topology.*

Here we remark that since S is a separable metric space then by Proposition 2.29 the Borel σ-algebra on $S \times S$ is the same as the product σ-algebra. Hence, we note that lower semicontinuous pseudometrics in $\mathfrak{lsc}_\mathfrak{m}$ are product measurable with respect to the unique σ-algebra on $S \times S$. Moreover, we have:

Proposition 3.6. *The spaces $\mathfrak{met}$ and $\mathfrak{lsc}_\mathfrak{m}$ are complete metric spaces when endowed with the metric induced by the uniform norm.*

Proposition 3.6 follows immediately by first noting that the set of bounded real-valued functions on $S \times S$ with the uniform norm metric is a complete metric space, and that $\mathfrak{met}$ and $\mathfrak{lsc}_\mathfrak{m}$ are closed subsets of this space.

Thus, once more we have a rich structure on our space of pseudometrics, admitting the use of important fixed point theorems, provided we construct an appropriate map on $\mathfrak{lsc}_\mathfrak{m}$. Doing so requires the use of a suitable probability metric; in light of the definition of bisimilarity, the importance of using the Kantorovich distance is made evident in the following lemma. Insofar as we know, this is an original result.

Lemma 3.7. *Let $h \in \mathfrak{lsc}_\mathfrak{m}$ as defined in Definition 3.5 and let $Rel(h)$ be the kernel of h as in Definition 2.9. Then $T_K(h)(P,Q) = 0$ if and only if $P(X) = Q(X)$, for every $X \in \Sigma(Rel(h))$.*

Proof. $\Leftarrow$ Fix $\epsilon > 0$ and let $f \in Lip(h)$ such that $T_K(h)(P,Q) < P(f) - Q(f) + \epsilon$. WLOG $f \geq 0$. Choose ψ a simple approximation (see Theorem 2.34) to f so that $T_K(h)(P,Q) < P(\psi) - Q(\psi) + 2\epsilon$. Let $\psi(S) = \{c_1, \ldots, c_k\}$ where the c_i are distinct, $E_i = \psi^{-1}(\{c_i\})$, and $R = Rel(h)$. Then each E_i is R-closed, for if $y \in R(E_i)$ then there is some $x \in E_i$ such that $h(x,y) = 0$. So $f(x) = f(y)$ and therefore, $\psi(x) = \psi(y)$. So $y \in E_i$. So by assumption $P(\psi) - Q(\psi) = \sum c_i P(E_i) - \sum c_i Q(E_i) = 0$. Thus, $T_K(h)(P,Q) = 0$.

$\Rightarrow$ Let $X \in \Sigma(R)$. Let $Y \subseteq X$ be compact. Define $f(x) = \inf_{y \in Y} h(x,y)$. Since a lower semicontinuous function has a minimum on a compact set, we may write $f(x) = \min_{y \in Y} h(x,y)$. In fact, f is itself lower semicontinuous by Theorem 2.23. Since f is measurable, $R(Y) = f^{-1}(\{0\}) \in \Sigma(R)$. By Theorem 2.33 and since S is a Polish metric space, P is tight, and hence $P(X) = \sup P(Y)$ where the supremum is taken over all compact $Y \subseteq X$. However, $Y \subseteq X$ implies $Y \subseteq R(Y) \subseteq R(X) = X$. Since $R(Y)$ is measurable, we have $P(X) = \sup P(R(Y))$. Similarly, $Q(X) = \sup Q(R(Y))$. Define $g_n = \max(0, 1 - nf)$ for $n \in \mathbb{N}$. Then the sequence $(g_n)_{n \in \mathbb{N}}$ decreases to the indicator function on $R(Y)$. Also, for each $n \in \mathbb{N}$, $g_n/n \in Lip(h)$. So by assumption $P(g_n/n) = Q(g_n/n)$ for every $n \in \mathbb{N}$. Multiplying by n and taking the limit as n tends to infinity gives $P(R(Y)) = Q(R(Y))$. ◻

The next result, which is original, essentially tells us that given the continuity assumptions on the MDP parameters, the limit of a sequence of pairs of bisimilar states is itself a pair of bisimilar states. First we need the following definitions:

Definition 3.8. *Let S be a Polish space. Then we define $\mathfrak{Equ}$ to be the set of equivlance relations on S equipped with subset ordering. We define $\mathfrak{Clo}_\mathfrak{E}$ to be the set of equivalence relations on S that are closed subsets of $S \times S$ endowed with the product topology.*

Proposition 3.9. *The sets $\mathfrak{Equ}$ and $\mathfrak{Clo}_\mathfrak{E}$ are complete lattices when equipped with the subset ordering.*

Clearly when we equip each set with the subset ordering, we obtain partial orders. The greatest lower bound of a set of equivalence relations is simply their intersection. Moreover, an arbitrary intersection of closed sets is closed. Hence, both spaces are complete lattices. Note that existence of least upper bounds follows from that of greatest lower bounds: least upper bounds are obtained as greatest lower bounds on the set of upper bounds.

Theorem 3.10. *Let (S, Σ, A, P, r) be an MDP satisfying the conditions of Definition 3.1. Then bisimilarity is a closed subset of $S \times S$*

Proof. Define $\mathcal{F} : \mathfrak{Equ} \to \mathfrak{Equ}$ by

$$s\mathcal{F}(R)s' \Leftrightarrow \text{ for every } a \in A,\ r^a_s = r^a_{s'} \text{ and for every } X \in \Sigma(R),\ P^a_s(X) = P^a_{s'}(X).$$

Then the greatest fixed point of $\mathcal{F}$ is bisimilarity.

That $\mathcal{F}(E)$ is an equivalence relation for a given E is obvious. That $\mathcal{F}$ has any fixed points at all is a consequence of the Knaster-Tarski Theorem, Theorem 2.25. Next, simply note that the fixed points of $\mathcal{F}$ are precisely the bisimulation relations. So the greatest fixed point is contained in bisimilarity, and since every bisimulation relation is contained in the greatest fixed point, so is bisimilarity.

We first claim that $\mathcal{F}$ maps $\mathfrak{Clo}_\mathfrak{E}$ to $\mathfrak{Clo}_\mathfrak{E}$. To see that $\mathcal{F}(E)$ is closed, let $(x_n, y_n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{F}(E)$ converging to some pair of states (x, y). Let $a \in A$. By the definition of $\mathcal{F}(E)$,

$r^a_{x_n} = r^a_{y_n}$ for every n. Since the reward function is continuous, taking the limit as n tends to infinity yields $r^a_x = r^a_y$. Next, let ρ_E be the discrete pseudometric assigning distance 1 to two points if and only if they are *not* related by E. Since E is closed, ρ_E is lower semicontinuous. So the Kantorovich metric, $T_K(\rho_E)$ is well-defined. Now we can invoke the leftmost equality in equation 2.5 to obtain that the map $(s, s') \mapsto T_K(\rho_E)(P^a_s, P^a_{s'})$ is lower semicontinuous; for since P^a_s is continuous with respect to the topology of weak convergence, $P^a_s(f)$ is continuous in the usual sense for every bounded continuous f in $Lip(\rho_E)$. So $P^a_s(f) - P^a_{s'}(f)$ is continuous on $S \times S$, and hence, lower semicontinuous. Finally, taking the supremum over every f yields that the map taking a pair of states to its Kantorovich distance with respect to ρ_E is lower semicontinuous. Let X be an E-closed measurable set. Then by definition of $\mathcal{F}(E)$, $P^a_{x_n}(X) = P^a_{y_n}(X)$, which by Lemma 3.7 means $T_K(\rho_E)(P^a_{x_n}, P^a_{y_n}) = 0$ for every n. Since $T_K(\rho_E)(P^a_s, P^a_{s'})$ is lower semicontinuous, $T_K(\rho_E)(P^a_x, P^a_y) = 0$. Again using Lemma 3.7, $P^a_x(X) = P^a_y(X)$. So (x, y) belongs to $\mathcal{F}(E)$, that is, $\mathcal{F}(E)$ is closed.

Now let $\sim_{\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}}$ be the least upper bound of bisimilarity in $\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}$. By monotonicity, we have $\sim\, = \mathcal{F}(\sim) \subseteq \mathcal{F}(\sim_{\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}})$. So $\sim_{\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}} \subseteq\, \mathcal{F}(\sim_{\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}})$, that is, $\sim_{\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}}$ is a postfixed point of $\mathcal{F}$; but then $\sim_{\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}} \subseteq\, \sim$, the latter being the greatest postfixed point.

Therefore, $\sim\, = \,\sim_{\mathcal{C}\mathfrak{lo}_{\mathfrak{E}}}$, that is, bisimilarity is closed. □

Definition 3.11. *A pseudometric ρ on the states of an MDP is a* bisimulation metric *if it satisfies $\rho(s, s') = 0 \iff s \sim s'$.*

All of the preceding theory comes together in the following crucial result. It is worth noting that our presentation is a significant extension of the work carried out by [58, 59] in their work on bisimulation metrics for labelled Markov processes.

Theorem 3.12. *Let $M = (S, \Sigma, A, P, r)$ be an MDP satisfying the conditions in Definition 3.1, $c \in (0, 1)$ be a metric discount factor, and $\mathfrak{lsc}_{\mathfrak{m}}$ be as in Definiton 3.5. Define $F : \mathfrak{lsc}_{\mathfrak{m}} \to \mathfrak{lsc}_{\mathfrak{m}}$ by*

$$F(h)(s, s') = \max_{a \in A}((1 - c)|r^a_s - r^a_{s'}| + cT_K(h)(P^a_s, P^a_{s'}))$$

Then

1. *F has a unique fixed point $\rho^* : S \times S \to [0, 1]$,*
2. *ρ^* is a bisimulation metric,*
3. *for any $h_0 \in \mathfrak{lsc}_{\mathfrak{m}}$, $\lim_{n\to\infty} F^n(h_0) = \rho^*$,*
4. *ρ^* is continuous on $S \times S$,*
5. *ρ^* is continuous in r and P, and*
6. *ρ^* scales with rewards, that is, if MDP $M' = (S, \Sigma, A, P, k \cdot r)$ for some $k \in [0, 1]$ then $\rho^*_{M'} = k \cdot \rho^*_M$.*

3.1. Proof of Theorem 3.12. The rest of this subsection will be dedicated to proving Theorem 3.12; however, let us first make a few remarks. The first three properties of the theorem tell us that a quantitative notion of bisimilarity exists, and that it can be approximated. The continuity results tell us that we only need to know the metric on a dense subset, and that distances are insensitive to perturbations in the MDP parameters. The last property is not surprising, and reflects the fact that the actual numbers are not as important as the qualitative structure arising from the metric. The topological or even *uniform* structures - see for example [20] - yield the same distinguishing information with respect to bisimilarity; our specific choice of pseudometric here is influenced by Theorem 3.20.

Lemma 3.13. *Let (S, Σ, A, P, r) be an MDP satisfying the conditions of Definition 3.1, $c \in (0, 1)$ be a metric discount factor, and F be as in the statement of Theorem 3.12. Then F has a unique* *fixed point* $\rho^* : S \times S \to [0,1]$*, such that for any* $h_0 \in \mathfrak{lsc}_{\mathfrak{m}}$*,*

$$\|\rho^* - F^n(h_0)\| \le \frac{c^n}{1-c}\|F(h_0) - h_0\|.$$

Proof.

We first need to ensure that F maps $\mathfrak{lsc}_{\mathfrak{m}}$ to $\mathfrak{lsc}_{\mathfrak{m}}$. Let h be a member of $\mathfrak{lsc}_{\mathfrak{m}}$. As in the proof of Theorem 3.10, we first note that for each action $a \in A$, the map taking (s, s') to $T_K(h)(P^a_s, P^a_{s'})$ is lower semicontinuous, as is the map taking (s, s') to $|r^a_s - r^a_{s'}|$. It follows that $F(h)$ is lower semicontinuous, since the sum of lower semicontinuous functions is lower semicontinuous and the maximum of lower semicontinuous functions is again lower semicontinuous.

Thus we obtain the result as a simple application of the Banach Fixed Point Theorem, Theorem 2.26, since $\mathfrak{lsc}_{\mathfrak{m}}$ is a complete metric space (Proposition 3.6). Here we use the dual minimization form of $T_K(\cdot)$, as given in equation 2.5. Note that for every $h, h' \in \mathfrak{lsc}_{\mathfrak{m}}$, and for every $s, s' \in S$,

$$\begin{aligned}
F(h)(s,s') - F(h')(s,s') &\le c \max_{a\in A}(T_K(h)(P^a_s, P^a_{s'}) - T_K(h')(P^a_s, P^a_{s'})) \\
&\le c \max_{a\in A}(T_K(h - h' + h')(P^a_s, P^a_{s'}) - T_K(h')(P^a_s, P^a_{s'})) \\
&\le c \max_{a\in A}(T_K(\|h - h'\| + h')(P^a_s, P^a_{s'}) - T_K(h')(P^a_s, P^a_{s'})) \\
&\le c \max_{a\in A}(\|h - h'\| + T_K(h')(P^a_s, P^a_{s'}) - T_K(h')(P^a_s, P^a_{s'})) \\
&\le c\|h - h'\|
\end{aligned}$$

In the third inequality, we have used monotonicity of the minimization form of $T_K(\cdot)(P^a_s, P^a_{s'})$ with respect to the cost function.

Thus, $\|F(h) - F(h')\| \le c\|h - h'\|$, so that F is a contraction mapping and has a unique fixed point ρ^*.

Note that for any $s, s' \in S$,

$$\begin{aligned}
\rho^*(s,s') = F(\rho^*)(s,s') &= \max_{a\in A}((1-c)|r^a_s - r^a_{s'}| + cT_K(\rho^*)(P^a_s, P^a_{s'})) \\
&\le \max_{a\in A}((1-c)\cdot 1 + cT_K(\|\rho^*\|)(P^a_s, P^a_{s'})) \\
&\le \max_{a\in A}((1-c) + c\|\rho^*\|) \\
&\le (1-c) + c\|\rho^*\|
\end{aligned}$$

whence it follows that $\|\rho^*\| \le (1-c) + c\|\rho^*\|$ and we conclude $\|\rho^*\| \le 1$. □

The following is an original continuity result.

Lemma 3.14. *Let* (S, Σ, A, P, r) *be an MDP satisfying the conditions of Definition 3.1 and let* ρ^* *be the pseudometric given by Lemma 3.13 with metric discount factor* $c \in (0,1)$*. Then* ρ^* *is a continuous function on* $S \times S$*.*

Proof. Since the set of bounded continuous pseudometrics on S is a closed subset of $\mathfrak{lsc}_{\mathfrak{m}}$, we need only show that F maps it to itself. So let ρ be a bounded continuous pseudometric on S. Let $a \in A$. Then continuity of r on S implies $|r^a_x - r^a_y|$ is continuous on $S \times S$. For the continuity of $T_K(\rho)(P^a_x, P^a_y)$, we appeal to Theorem 2.35. This theorem implies that $T_K(\rho)$ metrizes the topology of weak convergence, provided $Lip(\rho, C_b(S))$ is equicontinuous and uniformly bounded. Here we are using the leftmost equality in Theorem 2.39. Since ρ is bounded, $Lip(\rho, C_b(S))$ is uniformly bounded, as each member f of $Lip(\rho, C_b(S))$ maps to the interval $[0, \|\rho\|]$. As for equicontinuity at a point x, let $\epsilon > 0$. Continuity of the function $\rho(x, \cdot)$ implies that there is a neighborhood N_x of x such that for every y in N_x, $\rho(x, y) = |\rho(x, y) - \rho(x, x)| < \epsilon$. Then for any $f \in Lip(\rho, C_b(S))$, $|f(x) - f(y)| \leq \rho(x, y) < \epsilon$. Thus, $Lip(\rho, C_b(S)))$ is equicontinuous. Since

$$|T_K(\rho)(P^a_x, P^a_y) - T_K(\rho)(P^a_{x_n}, P^a_{y_n})| \leq T_K(\rho)(P^a_x, P^a_{x_n}) + T_K(\rho)(P^a_y, P^a_{y_n})$$

we have that for any $(x_n, y_n)_{n \in \mathbb{N}}$ converging to (x, y), $T_K(\rho)(P^a_{x_n}, P^a_{y_n})$ converges to $T_K(\rho)(P^a_x, P^a_y)$. Thus, continuity of $F(\rho)$ is immediate. □

As an immediate consequence of Lemma 3.14, we have the following:

Corollary 3.15. *Let (S, Σ, A, P, r) be an MDP satisfying the conditions of Definition 3.1 and let ρ^* be the pseudometric given by Lemma 3.13 with metric discount factor $c \in (0, 1)$. Then the topology induced by ρ^* on S is coarser than the original.*

Next we show that we have indeed quantitatively captured bisimilarity. The proof of this result is original.

Lemma 3.16. *Let (S, Σ, A, P, r) be an MDP satisfying the conditions of Definition 3.1 and let ρ^* be the pseudometric given by Lemma 3.13 with metric discount factor $c \in (0, 1)$. Then ρ^* is a bisimulation metric*

Proof. It follows from Lemma 3.7 that for any h in $\mathfrak{lsc}_{\mathfrak{m}}$, $Rel(F(h)) = \mathcal{F}(Rel(h))$. Thus, $Rel(\rho^*) = \mathcal{F}(Rel(\rho^*))$ is a fixed point and so is contained in bisimilarity. For the other direction, we consider the discrete bisimilarity pseudometric that assigns distance 1 to pairs of non-bisimilar states; call it ρ. Since bisimilarity is closed (Theorem 3.10), ρ is lower semicontinuous. So $\sim\, = \mathcal{F}(\sim)\; = \mathcal{F}(Rel(\rho)) = Rel(F(\rho))$, which implies $F(\rho) \leq \rho$. Since F is monotone, iterating F and taking limits yields $\rho^* \leq \rho$, whence it follows that $Rel(\rho^*)$ contains bisimilarity. □

Before moving on, let us give meaning to the iterates $\{F^n(\bot) : n \in \mathbb{N}\}$. Define inductively $\sim_0 = S \times S$, and $\sim_{n+1} = \mathcal{F}(\sim_n)$. Finally, let $\sim_\omega = \cap_n \sim_n$ represent the limit of this sequence.

The best way to view this is once more in terms of "information resolution". At first, we know nothing; this is represented by the relation that equates all states, $\sim_0$. Applying $\mathcal{F}$ corresponds to a one-step lookahead refinement, and similarly for n steps. Our intuition naturally tells us that in the limit, we should have a "strong matching", that is, bisimilarity; however, it is not immediately clear that this is so. Not surprisingly, a proof once more makes itself evident through the use of metrics.

Simply note that by induction $Rel(F^n(\bot)) = \sim_n$ (here, we are once again using the fact that $Rel(F(h)) = \mathcal{F}(Rel(h))$). Since it is easily seen that $\cap_n Rel(F^n(\bot)) = Rel(\sup_n F^n(\bot))$ and $\sup_n F^n(\bot) = \rho^*$, we have $\sim_\omega = Rel(\rho^*)$, which is bisimilarity.

Thus, the nth iterate corresponds to an n-step approximation to bisimilarity. Let us note that we now have three equivalent formulations of bisimilarity, making this more in line with the traditional presentation of bisimilarity for labeled nondeterministic transition systems: as a maximal relation, as a greatest fixed point, and as an intersection of an infinite family of equivalence relations [44].

Lemma 3.17. *Let $M = (S, \Sigma, A, P, r)$ and $M' = (S, \Sigma, A, P, r')$ be two MDPs satisfying the conditions of Definition 3.1 and let ρ^*_M and $\rho^*_{M'}$ be their respective pseudometrics given by Lemma 3.13 with common metric discount factor $c \in (0, 1)$. If $r' = k \cdot r$ for some scalar $k > 0$, then $\rho^*_{M'} = k \cdot \rho^*_M$.*

Proof. It is not hard to see that $k \cdot \rho^*_M$ is a solution to the fixed point equation for M'; thus, the result follows by uniqueness. □

The following original result shows that, by contrast with bisimilarity, the bisimilarity distances vary smoothly with the MDP parameters.

Lemma 3.18. *Let $M = (S, \Sigma, A, r, P)$ and $\widehat{M} = (S, \Sigma, A, \hat{r}, Q)$ be two MDPs with common state and action spaces, and satisfying the conditions of Definition 3.1. Let ρ and $\hat{\rho}$ be the corresponding bisimulation metrics given by Lemma 3.13 with discount factor $c \in (0, 1)$. Then*

$$\|\rho - \hat{\rho}\| \leq 2\|r - \hat{r}\| + \frac{2c}{(1-c)} \sup_{a,s} TV(P^a_s, Q^a_s),$$

where TV is the total variation probability metric, as defined by

$$TV(P, Q) = \sup_{X \in \Sigma} |P(X) - Q(X)|.$$

Proof. Let d be the discrete pseudometric that assigns distance 1 to all pairs of non-equal states. Using the triangle inequality along with the fact that $Lip(h)$ is contained in $Lip(d)$ for $h \in \{\rho, \hat{\rho}\}$, we first obtain:

$$\begin{aligned} T_K(h)(P^a_x, P^a_y) - T_K(h)(Q^a_x, Q^a_y) &\leq T_K(h)(P^a_x, Q^a_x) + T_K(h)(P^a_y, Q^a_y) \\ &\leq T_K(d)(P^a_x, Q^a_x) + T_K(d)(P^a_y, Q^a_y) \\ &\leq TV(P^a_x, Q^a_x) + TV(P^a_y, Q^a_y) \end{aligned} \tag{3.1}$$

Here we have used the fact that $T_K(d) = TV$ [60]. Next, using monotonicity of $T_K(\cdot)$ with respect to the cost function, we have

$$\begin{aligned} T_K(\rho)(P^a_x, P^a_y) - T_K(\hat{\rho})(P^a_x, P^a_y) &= T_K(\rho - \hat{\rho} + \hat{\rho})(P^a_x, P^a_y) - T_K(\hat{\rho})(P^a_x, P^a_y) \\ &\leq T_K(\|\rho - \hat{\rho}\| + \hat{\rho})(P^a_x, P^a_y) - T_K(\hat{\rho})(P^a_x, P^a_y) \\ &= \inf_{\lambda \in \Lambda(P^a_x, P^a_y)} \int_{S \times S} (\|\rho - \hat{\rho}\| + \hat{\rho}) d\lambda - T_K(\hat{\rho})(P^a_x, P^a_y) \\ &= \inf_{\lambda \in \Lambda(P^a_x, P^a_y)} (\|\rho - \hat{\rho}\| + \int_{S \times S} \hat{\rho} d\lambda) - T_K(\hat{\rho})(P^a_x, P^a_y) \\ &= \|\rho - \hat{\rho}\| + T_K(\hat{\rho})(P^a_x, P^a_y) - T_K(\hat{\rho})(P^a_x, P^a_y) \\ &\leq \|\rho - \hat{\rho}\| \end{aligned} \tag{3.2}$$

Finally

$$\begin{aligned} &\rho(x, y) - \hat{\rho}(x, y) \\ &\leq \max_{a \in A}((1-c)|r^a_x - r^a_y| + cT_K(\rho)(P^a_x, P^a_y)) - \max_{a \in A}((1-c)|\hat{r}^a_x - \hat{r}^a_y| + cT_K(\hat{\rho})(Q^a_x, Q^a_y)) \\ &\leq \max_{a \in A}((1-c)(|r^a_x - r^a_y| - |\hat{r}^a_x - \hat{r}^a_y|) + c(T_K(\rho)(P^a_x, P^a_y) - T_K(\hat{\rho})(Q^a_x, Q^a_y))) \\ &\leq \max_{a \in A}((1-c)(|(r^a_x - r^a_y) - (\hat{r}^a_x - \hat{r}^a_y)|) \\ &\qquad + c(T_K(\rho)(P^a_x, P^a_y) - T_K(\hat{\rho})(P^a_x, P^a_y)) + c(T_K(\hat{\rho})(P^a_x, P^a_y) - T_K(\hat{\rho})(Q^a_x, Q^a_y))) \\ &\leq \max_{a \in A}((1-c)(|r^a_x - \hat{r}^a_x| + |r^a_y - \hat{r}^a_y|) + c\|\rho - \hat{\rho}\| + 2c \sup_s TV(P^a_s, Q^a_s))) \\ &\leq \max_{a \in A}(2(1-c)\|r^a - \hat{r}^a\| + c\|\rho - \hat{\rho}\| + 2c \sup_s TV(P^a_s, Q^a_s)) \\ &\leq 2(1-c)\|r - \hat{r}\| + c\|\rho - \hat{\rho}\| + 2c \sup_{a,s} TV(P^a_s, Q^a_s))) \end{aligned}$$

□

Now suppose we are given an MDP and another MDP is alleged to be a good approximation. We would ideally like to measure the distance between a state in the original model and its equivalent state in the approximation using a bisimulation metric. The next results tells us that we can do so in a well-defined manner.

Theorem 3.19. *Let $M_1 = (S_1, \Sigma_1, A, r_1, P_1)$ and $M_1 = (S_2, \Sigma_2, A, r_2, P_2)$ be two MDPs that satisfy the conditions of Definition 3.1. Suppose S_1 and S_2 are disjoint. Define the disjoint union of M_1 and M_2 to be $M = (S, \Sigma, A, r, P)$ where*

- *$S = S_1 \sqcup S_2$ is the disjoint union of S_1 and S_2,*
- *Σ is the Borel σ-algebra on S,*
- *$r : S \times A \to [0, 1]$ is defined by $r(x, a) = r_i(x, a)$ if $x \in S_i$ for $i = 1, 2$, and*
- *$P : S \times A \times \Sigma \to [0, 1]$ is defined by $P(x, a, B) = P_i(x, a, B \cap S_i)$ if $x \in S_i$ for $i = 1, 2$.*

Then M is an MDP satisfying the conditions of Defintion 3.1. Moreover, if $\rho^*_{M_1}$, $\rho^*_{M_2}$, and ρ^*_M are the bisimulation metrics guaranteed by Lemma 3.13 with metric discount factor $c \in (0, 1)$ then the restriction $\rho^*_M \restriction_{M_i} = \rho^*_{M_i}$ for $i = 1, 2$.

Proof. Let d_1 and d_2 be metrics inducing the respective topologies of S_1 and S_2 as Polish spaces. Endow S with the metric d defined by $d(x, y) = d_i(x, y)$ if $x, y \in S_i$ and 1 otherwise. Then it is not hard to see that (S, d) is Polish metric space, and S is Polish. So it makes sense to speak of its Borel σ-algebra Σ.

Let $(x_n)_{n\in\mathbb{N}}$ be a convergent sequence in S converging to some point $x \in S$. By Definition 2.10, we can choose $N \in \mathbb{N}$ such that for each $n \geq N$, $d(x_n, x) < \frac{1}{2}$. So $(x_n)_{n\geq N}$ and x must belong solely to one of S_1 and S_2 and continuity of r and P follow immediately from the sequential version of continuity in each of the spaces S_1 and S_2. Let us outline the argument for P.

First note that P is a well-defined stochastic transition kernel follows from P_1 and P_2 being stochastic transition kernels. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in S converging to a point $x \in S$. Without loss of generality, we can assume that these belong entirely to S_1. Let f be a bounded continuous function on S. Its restriction $f \restriction_{S_1}$ is easily seen to be a bounded continuous function on S_1 (again, use the sequential version of continuity). Then

$$\int_S f dP^a_{x_n} = \int_{S_1} f \restriction_{S_1} dP^a_{1,x_n} + \int_{S_2} f \restriction_{S_2} dP^a_{2,x_n} = \int_{S_1} f \restriction_{S_1} dP^a_{1,x_n} + 0 = \int_{S_1} f \restriction_{S_1} dP^a_{1,x_n}$$

Similarly, $\int_S f dP^a_x = \int_{S_1} f \restriction_{S_1} dP^a_{1,x}$. Hence,

$$\lim_{n\to\infty} \int_S f dP^a_{x_n} = \lim_{n\to\infty} \int_{S_1} f \restriction_{S_1} dP^a_{1,x_n} = \int_{S_1} f \restriction_{S_1} dP^a_{1,x} = \int_S f dP^a_x$$

Therefore M satisfies the conditions of Defintion 3.1. By Theorem 3.12, bisimulation metrics $\rho^*_{M_1}$,

$\rho^*_{M_2}$, and ρ^*_M exist and are unique. Let us show that $\rho^*_M \restriction_{M_1} = \rho^*_{M_1}$. First let $x, y \in S_1$. Then

$$\begin{aligned}
T_K(\rho^*_M)(P^a_x, P^a_y) &= \sup_{f \in Lip(\rho^*_M)} \left(\int_S f dP^a_x - \int_S f dP^a_y\right) \\
&= \sup_{f \in Lip(\rho^*_M)} \left(\int_{S_1} f \restriction_{S_1} dP^a_{1,x} + \int_{S_2} f \restriction_{S_2} dP^a_{1,x} - \int_{S_1} f \restriction_{S_1} dP^a_{1,y} - \int_{S_2} f \restriction_{S_2} dP^a_{1,y}\right) \\
&= \sup_{f \in Lip(\rho^*_M)} \left(\int_{S_1} f \restriction_{S_1} dP^a_{1,x} + 0 - \int_{S_1} f \restriction_{S_1} dP^a_{1,y} - 0\right) \\
&= \sup_{f \in Lip(\rho^*_M)} \left(\int_{S_1} f \restriction_{S_1} dP^a_{1,x} - \int_{S_1} f \restriction_{S_1} dP^a_{1,y}\right) \\
&\leq T_K(\rho^*_M \restriction_{M_1})(P^a_{1,x}, P^a_{1,y})
\end{aligned}$$

Now fix $\epsilon > 0$. Then there exists an $f \in Lip(\rho^*_M \restriction_{M_1})$ such that

$$T_K(\rho^*_M \restriction_{M_1})(P^a_{1,x}, P^a_{1,y}) - \epsilon < \int_{S_1} f dP^a_{1,x} - \int_{S_1} f dP^a_{1,y}$$

Define $g : S \to \mathbb{R}$ by $g(z) = \inf_{s \in S_1}(f(s) + \rho^*_M(z, s))$. Then $g \in Lip(\rho^*_M)$ and $g \restriction_{S_1} = f$. Note by Lemma 3.14 that since g is Lipschitz continuous with respect to ρ^*_M then g is in fact continuous on S with its given topology, and hence measurable. Next,

$$\begin{aligned}
T_K(\rho^*_M \restriction_{M_1})(P^a_{1,x}, P^a_{1,y}) - \epsilon &< \int_{S_1} f dP^a_{1,x} - \int_{S_1} f dP^a_{1,y} \\
&= \int_{S_1} g \restriction_{S_1} dP^a_{1,x} + 0 - \int_{S_1} g \restriction_{S_1} dP^a_{1,y} - 0 \\
&= \int_{S_1} g \restriction_{S_1} dP^a_{1,x} + \int_{S_2} g \restriction_{S_2} dP^a_{1,x} - \int_{S_1} g \restriction_{S_1} dP^a_{1,y} - \int_{S_2} g \restriction_{S_2} dP^a_{1,y} \\
&= \int_S g dP^a_x - \int_S g dP^a_y \\
&\leq T_K(\rho^*_M)(P^a_x, P^a_y)
\end{aligned}$$

Since ϵ is arbitrary, we conclude $T_K(\rho^*_M)(P^a_x, P^a_y) = T_K(\rho^*_M \restriction_{M_1})(P^a_{1,x}, P^a_{1,y})$. Therefore, for any $s, s' \in S_1$,

$$\begin{aligned}
\rho^*_M(s, s') &= \max_{a \in A}((1-c)|r^a_s - r^a_{s'}| + cT_K(\rho^*_M)(P^a_s, P^a_{s'})) \\
&= \max_{a \in A}((1-c)|r^a_{1,s} - r^a_{1,s'}| + cT_K(\rho^*_M \restriction_{M_1})(P^a_{1,s}, P^a_{1,s'})) \\
&= F_M(\rho^*_M \restriction_{M_1})(s, s')
\end{aligned}$$

where F_{M_1} is the fixed-point operator for $\rho^*_{M_1}$. Thus, it follows that $F_{M_1}(\rho^*_M \restriction_{M_1}) = \rho^*_M \restriction_{M_1}$. By uniqueness, we conclude $\rho^*_M \restriction_{M_1} = \rho^*_{M_1}$. □

Thus, existence of the state-similarity metrics for a continuous MDP is established, along with several important properties. However, as in the finite case, perhaps the most important property of the metrics is showing that similar states have similar optimal values, and that this relation varies smoothly with similarity. We must emphasize that in contrast with the work on LMPs, where the underlying motivation has been to analyze the validity of testing properties expressed in a modal logic on similar systems, a primary focus here is in analyzing the validity of computing optimal values (and hence, optimal policies) on similar MDPs.

3.2. Value Function Bounds. In moving to continuous state spaces, we must address the validity of the continuous analog of the optimality equations:

$$V^*(s) = \max_{a \in A} (r_s^a + \gamma P_s^a(V^*)) \text{ for each } s \in S.$$

In general, such a V^* need not exist. Even if it does, there may not be a well-behaved, that is to say measurable, policy that is captured by it. Fortunately, there are several mild restrictions under which this is not the case; and in fact, Theorem 6.2.12. of [53] states that the optimality equations are valid provided the state space is Polish and the reward function is uniformly bounded, as is indeed the case here. Just as before, the optimal value function V^* can be expressed as the limit of a sequence of iterates V^n; we can use these to show that the optimal value function is continuous with respect to the state-similarity metrics.

Theorem 3.20. *Let $M = (S, \Sigma, A, P, r)$ be an MDP satisfying the conditions of Definition 3.1, ρ^* be the pseudometric given by Theorem 3.12 with metric discount factor $c \in (0, 1)$, and V^* be the optimal value function for M with discount factor $\gamma \in [0, 1)$. Suppose $\gamma \leq c$. Then V^* is Lipschitz continuous with respect to ρ^* with Lipschitz constant $\frac{1}{1-c}$, i.e.*

$$|V^*(s) - V^*(s')| \leq \frac{1}{1-c} \rho^*(s, s').$$

Proof. Each iterate V^n is continuous, and so each $|V^n(s) - V^n(s')|$ belongs to $\mathfrak{lsc}_\mathfrak{m}$. The result now follows by induction and taking limits. □

We can use this result to relate the optimal values of a state and its representation in an approximant by considering the original model and its approximant as one MDP. More directly, we can use the distances themselves for aggregation with error bounds. Let us consider a simple illustration, first presented in [28], of metric-based reasoning: let $S = [0, 1]$ with the usual Borel σ-algebra, $A = \{a, b\}$, $r_s^a = 1 - s$, $r_s^b = s$, P_s^a be uniform on S, and P_s^b the point mass at s. Clearly, these MDP parameters satisfy the required assumptions.

Given any $c \in (0, 1)$, we claim $\rho^*(x, y) = |x - y|$. Define h by $h(x, y) = |x - y|$, and note that $T_K(h)(P_x^a, P_y^a) = 0$ and $T_K(h)(P_x^b, P_y^b) = h(x, y)$. Thus, $F(h)(x, y) = \max((1 - c)|x - y| + c \cdot 0, (1 - c)|x - y| + c \cdot h(x, y)) = (1 - c)h(x, y) + c \cdot h(x, y) = h(x, y)$. By uniqueness, $\rho^* = h$ as was to be shown.

Now consider the following approximation. Given $\epsilon > 0$, choose n large enough so that $\frac{1}{n} < (1 - c)\epsilon$. Partition S as $B_k = [\frac{k}{n}, \frac{k+1}{n})$, $B_{n-1} = [\frac{n-1}{n}, 1]$, for $k = 0, 1, 2, \ldots, n - 2$. Note that the diameter of each B_k with respect to ρ^*, $\mathrm{diam}_{\rho^*} B_k$, is $\frac{1}{n} < (1 - c)\epsilon$. The n partitions will be the states of a finite MDP approximant. We obtain the rest of the parameters by averaging over the states in a partition. Thus, $r_{B_k}^a = 1 - \frac{2k+1}{2n}$, $r_{B_k}^b = \frac{2k+1}{2n}$, $P_{B_k, B_l}^a = \frac{1}{n}$, and $P_{B_k, B_l}^b = 1$ if $k = l$ and 0 otherwise.

Assume γ is given and choose $c = \gamma$. Note that for every x, y in B_k,

$$|V^*(x) - V^*(y)| \leq \frac{1}{1-c} \mathrm{diam}_{\rho^*} B_k \leq \epsilon.$$

Thus, we would expect that by averaging, and solving the finite MDP, $V^*(B_k)$ should differ by at most ϵ from $V^*(x)$, for any $x \in B_k$. In fact, in this case the value functions of the original MDP and of the finite approximant can be computed directly and we can verify this. For $x \in S$, B_k, we find:

$$V^*(x) = \begin{cases} 1 - x + \frac{\gamma}{2(1-\gamma)} & \text{if } 0 \leq x < \frac{1}{2} \\ \frac{x}{1-\gamma} & \text{if } \frac{1}{2} \leq x \leq 1 \end{cases}$$

$$V^*(B_k) = \begin{cases} 1 - \frac{2k+1}{2n} + \frac{\gamma}{2(1-\gamma)} & \text{if } 0 \leq k < \frac{n-1}{2} \\ \frac{\frac{2k+1}{2n}}{1-\gamma} & \text{if } \frac{n-1}{2} \leq k \leq n-1 \end{cases}$$

Therefore, for $x \in B_k$,

$$|V^*(x) - V^*(B_k)| \leq \frac{1}{1-\gamma}\left| x - \frac{2k+1}{2n} \right| \leq \frac{1}{1-c} \operatorname{diam}_{\rho^*} B_k \leq \epsilon.$$

In fact, we can somewhat generalize this result.

Theorem 3.21. *Let $M = (S, \Sigma, A, r, p)$ be an MDP satisfying the conditions of Definition 3.1. Let μ be a measure on S, and $\mathcal{P}$ a finite partition of S such that each equivalence in $\mathcal{P}$ has positive μ-measure. Let $[-] : S \to \mathcal{P}$ be the map that takes $s \in S$ to its equivalence class in $\mathcal{P}$. Define the μ-average finite MDP $M_\mathcal{P}$ by $(\mathcal{P}, A, r, P)$ where*

$$r_B^a = \frac{1}{\mu(B)} \int_{x \in B} r_x^a d\mu(x) \text{ and } P_{BB'}^a = \frac{1}{\mu(B)} \int_{x \in B} P_x^a(B') d\mu(x).$$

Let $\rho^*_{M \sqcup M_\mathcal{P}}$ be the bisimulation metric for the disjoint union of M and $M_\mathcal{P}$ with metric discount factor $c \in (0,1)$ as given by Theorem 3.19. Let V_M^* and $V_\mathcal{P}^*$ be the optimal value functions for M and $M_\mathcal{P}$ respectively with discount factor $\gamma \in [0,1)$. Suppose $\gamma \leq c$. Then for every $s \in S$,

$$(1-c) \cdot |V_M^*(s) - V_{M_\mathcal{P}}^*([s])| \leq \rho^*_{M \sqcup M_\mathcal{P}}(s, [s]) \leq \frac{1}{1-c} \sup_{y \in s} \frac{1}{\mu([y])} \int_{x \in [y]} \rho_M^*(y, x) d\mu(x). \qquad (3.3)$$

In other words, we can bound the distance between a state and its equivalence class by the maximum average distance between a state and all the other states in its equivalence class.

Proof. First note that since $\mathcal{P}$ is a finite set, we can endow it with the discrete metric assigning distance 1 to all pairs of non-equal states to make it a Polish space. Then $M_\mathcal{P}$ trivially satisfies the conditions of Definition 3.1. So by Theorem 3.19 the disjoint union of M and $M_\mathcal{P}$ exists, as does the bisimulation metric $\rho^*_{M \sqcup M_\mathcal{P}}$. Therefore, the lefthand equality in equation 3.3 follows from Theorem 3.20. Here we note that the value function is defined over the disjoint union MDP, but agrees on values restricted to the individual MDPs, just as is the case for the bisimulation metrics.

For the righthand equality, let $s \in S$ and $a \in A$. Let $\epsilon > 0$. Then there exists an f in

$Lip(\rho^*_{M\sqcup M_{\mathcal{P}}})$ such that

$$\begin{aligned}
T_K(\rho^*_{M\sqcup M_{\mathcal{P}}})(P^a_s, P^a_{[s]}) - \epsilon < \int_{S\sqcup\mathcal{P}} f dP^a_s - \int_{S\sqcup\mathcal{P}} f dP^a_{[s]} = \int_{x\in S} f(x)dP^a_s(x) - \int_{B\in\mathcal{P}} f(B)dP^a_{[s]}(B)\\
&= \int_{x\in S} f(x)dP^a_s(x) - \sum_{B\in\mathcal{P}} f(B)P^a_{[s]}(B)\\
&= \int_{x\in S} f(x)dP^a_s(x) - \frac{1}{\mu([s])}\int_{z\in[s]}(\int_{x\in S} f(x)dP^a_z(x))d\mu(z)\\
&\qquad + \frac{1}{\mu([s])}\int_{z\in[s]}(\int_{x\in S} f(x)dP^a_z(x))d\mu(z) - \sum_{B\in\mathcal{P}} f(B)\frac{1}{\mu([s])}\int_{z\in[s]} P^a_z(B)d\mu(z)\\
&= \frac{1}{\mu([s])}\int_{z\in[s]}(\int_{x\in S} f(x)dP^a_s(x) - \int_{x\in S} f(x)dP^a_z(x))d\mu(z)\\
&\qquad + \frac{1}{\mu([s])}\int_{z\in[s]}(\int_{x\in S} f(x)dP^a_z(x) - \sum_{B\in\mathcal{P}} f(B)P^a_z(B))d\mu(z)\\
&\le \frac{1}{\mu([s])}\int_{z\in[s]} T_K(\rho^*_M)(P^a_s, P^a_z)d\mu(z)\\
&\qquad + \frac{1}{\mu([s])}\int_{z\in[s]}(\int_{x\in S} f(x)dP^a_z(x) - \sum_{B\in\mathcal{P}}\int_{x\in B} f(B)dP^a_z(x))d\mu(z)\\
&\le \frac{1}{\mu([s])}\int_{z\in[s]} T_K(\rho^*_M)(P^a_s, P^a_z)d\mu(z)\\
&\qquad + \frac{1}{\mu([s])}\int_{z\in[s]}(\int_{x\in S} f(x)dP^a_z(x) - \int_{x\in S} f([x])dP^a_z(x))d\mu(z)\\
&\le \frac{1}{\mu([s])}\int_{z\in[s]} T_K(\rho^*_M)(P^a_s, P^a_z)d\mu(z) + \frac{1}{\mu([s])}\int_{z\in[s]}\int_{x\in S}(f(x) - f([x]))dP^a_z(x)d\mu(z)\\
&\le \frac{1}{\mu([s])}\int_{z\in[s]} T_K(\rho^*_M)(P^a_s, P^a_z)d\mu(z) + \frac{1}{\mu([s])}\int_{z\in[s]}\int_{x\in S}\rho^*_{M\sqcup M_{\mathcal{P}}}(x,[x])dP^a_z(x)d\mu(z)\\
&\le \frac{1}{\mu([s])}\int_{z\in[s]} T_K(\rho^*_M)(P^a_s, P^a_z)d\mu(z) + \sup_{x\in S}\rho^*_{M\sqcup M_{\mathcal{P}}}(x,[x])
\end{aligned}$$

Then

$$\begin{aligned}
\rho^*_{M\sqcup M_{\mathcal{P}}}(s,[s]) &= \max_{a\in A}((1-c)|r^a_s - r^a_{[s]}| + cT_K(\rho^*_{M\sqcup M_{\mathcal{P}}})(P^a_s, P^a_{[s]}))\\
&\le \max_{a\in A}((1-c)\frac{1}{\mu([s])}\int_{z\in[s]}|r^a_s - r^a_z|d\mu(z)\\
&\qquad + c(\frac{1}{\mu([s])}\int_{z\in[s]} T_K(\rho^*_M)(P^a_s, P^a_z)d\mu(z) + \sup_{x\in S}\rho^*_{M\sqcup M_{\mathcal{P}}}(x,[x])))\\
&\le \frac{1}{\mu([s])}\int_{z\in[s]}\max_{a\in A}((1-c)|r^a_s - r^a_z| + cT_K(\rho^*_M)(P^a_s, P^a_z))d\mu(z) + c\cdot\sup_{x\in S}\rho^*_{M\sqcup M_{\mathcal{P}}}(x,[x]))\\
&\le \frac{1}{\mu([s])}\int_{z\in[s]}\rho^*_M(s,z)d\mu(z) + c\cdot\sup_{x\in S}\rho^*_{M\sqcup M_{\mathcal{P}}}(x,[x]))
\end{aligned}$$

Taking the supremum on both sides with respect to s, we find

$$\sup_{s \in S} \rho^*_{M \sqcup M_{\mathcal{P}}}(s, [s]) \leq \frac{1}{1-c} \sup_{s \in s} \frac{1}{\mu([s])} \int_{z \in [s]} \rho^*_M(s, z) d\mu(z)$$

□

4. More Mathematical Review. When dealing with infinite state spaces in practice, we still need to find some finite structure with which to work; therefore, we restrict our attention here to those Markov decision processes whose state spaces are compact metric spaces. As we will be sampling in such spaces, we will also need some results on empirical processes. We present here some mathematical definitions and results for empirical processes valid on compact metric spaces. These can be found in [29] and [19].

4.1. A Compactness Theorem. THEOREM 4.1 (Arzelà-Ascoli). *Let X be a compact Hausdorff space and $C(X)$ be the space of continuous complex-valued functions on X. If $\mathcal{F}$ is an equicontinuous, pointwise bounded subset of $C(X)$, then $\mathcal{F}$ is totally bounded in the uniform metric, and the closure of $\mathcal{F}$ in $C(X)$ is compact.*

4.2. Weak Convergence and Empirical Processes. DEFINITION 4.2. *Let $n \in \mathbb{N}$ and let $(\Omega, \mathcal{A}, \mathbb{P})$ be an ambient probability space over which we sample n points $\{X_1, X_2, \ldots, X_n\}$ with values in (S, Σ) independently and with identical distribution μ; that is, each X_i is a measurable map from $(\Omega, \mathcal{A}, \mathbb{P})$ to (S, Σ) such that $\mathbb{P}(\{\omega \in \Omega : X_i(\omega) \in E\}) = \mathbb{P}(X_i^{-1}(E)) = \mu(E)$ for every E in Σ. Define the nth* empirical probability measure *μ_n of μ to be the average of the Dirac measures at each X_i; that is, $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.*

Each μ_n is in effect a random measure; that is, for each $\omega \in \Omega$, $\mu_n(\omega) := \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)}$ is a probability measure. Does this sequence of random probability measures $(\mu_n)_{n \in \mathbb{N}}$ converge?

DEFINITION 4.3. *Let $(Y_n)_{n \in \mathbb{N}}$ be a sequence of random variables and let Y be a random variable. Then $(Y_n)_{n \in \mathbb{N}}$ converges to Y* in probability, *if for every $\epsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}(\{\omega \in \Omega : |Y_n(\omega) - Y(\omega)| \geq \epsilon\}) = 0,$$

and almost surely *if*

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n \to \infty} Y_n(\omega) = Y(\omega)\}) = 1.$$

The Weak Law of Large Numbers (Strong Law of Large Numbers) tells us that for each real-valued bounded continuous f, the sequence of random variables $(\mu_n(f))_{n \in \mathbb{N}} = (\frac{1}{n}\sum_{i=1}^n f(X_i))_{n \in \mathbb{N}}$ converges to $\mu(f)$ in $\mathbb{P}$-probability ($\mathbb{P}$-almost surely). If the convergence was uniform over the set $\mathcal{F}$ of all bounded continuous functions, that is, if $\sup_{f \in \mathcal{F}} |\mu_n(f) - \mu(f)|$ converged to zero , then it would follow that the empirical measures themselves converged weakly. This turns out to be a useful property in itself. Let us note that the maps $\omega \mapsto \sup_{f \in \mathcal{F}} |\mu_n(\omega)(f) - \mu(f)|$ need not be measurable since they involve taking a supremum over the possibly uncountable collection $\mathcal{F}$. Thus, we will need to use the outer probability $\mathbb{P}^*$ when speaking of their convergence.

DEFINITION 4.4. *Let $\mathcal{F}$ be a class of integrable functions for probability measure μ. Then $\mathcal{F}$ is a* weak Glivenko-Cantelli class for μ *if $\sup_{f \in \mathcal{F}} |\mu_n(f) - \mu(f)|$ converges to zero in $\mathbb{P}^*$-outer probability. It is a* strong Glivenko-Cantelli class for μ *if convergence is $\mathbb{P}^*$-almost sure.*

DEFINITION 4.5. *If $\mathcal{F}$ is a Glivenko-Cantelli class for every probability measure on (S, Σ) then it is said to be a* universal Glivenko-Cantelli class. *Lastly, if the rate of $\mathbb{P}^*$-convergence can be made* *to be uniform over all μ, that is, if for every positive ϵ there exists a natural number N depending only on ϵ such that for every μ and every $n \geq N$, $\mathbb{P}^*(\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)| > \epsilon) < \epsilon$, then $\mathcal{F}$ is said to be a* a strong uniform Glivenko-Cantelli class.

An equivalent formulation says that $\mathcal{F}$ is a strong uniform Glivenko-Cantelli class if and only if

$$\text{for every } \epsilon > 0 \lim_{i\to\infty} \sup_{\mu} \mathbb{P}^*(\sup_{m\geq i} \sup_{f\in\mathcal{F}} |\mu(f) - \mu_m(f)| > \epsilon) = 0,$$

where the outermost supremum is taken over all probability measures on the state space.

The following definitions are needed to set up a Glivenko-Cantelli theorem, which will be used to establish that a certain class of functions, $Lip(\rho^*, C_b(S))$, is a strong uniform Glivenko-Cantelli class when S is a compact metric space.

DEFINITION 4.6. *A topological space is an* analytic space*, also known as a* Suslin space*, if it is the continuous image of a Polish space under a map between Polish spaces.*

DEFINITION 4.7. *Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathcal{F}$ a set. Then a real-valued function $X : (f, w) \mapsto X(f, w)$ will be called* image admissible Suslin via $(Y, \mathcal{S}, T)$ *if and only if $(Y, \mathcal{S})$ is a Suslin measurable space, T is a function from Y onto $\mathcal{F}$, and the map $(y, \omega) \mapsto X(T(y), \omega)$ is jointly measurable on $Y \times \Omega$.*

If $\mathcal{F}$ is a set of functions on Ω and $X(f, \omega) = f(\omega)$ is the evaluation map then $\mathcal{F}$ will be called image admissible Suslin *if X is image admissible Suslin via some $(Y, \mathcal{S}, T)$ as above.*

DEFINITION 4.8. *Let X be a set, $x = (x_1, \ldots, x_n) \in X^n$ for $n = 1, 2, \ldots$, and $\mathcal{F}$ be a family of real-vaued functions on X Define the pseudometric $e_{x,\infty}$ on $\mathcal{F}$ by*

$$e_{x,\infty}(f, g) = \max_{1\leq i\leq n} |f(x_i) - g(x_i)| \text{ for } f, g \in \mathcal{F}$$

Let $N(\epsilon, \mathcal{F}, e_{x,\infty})$ denote the ϵ-covering number of $(\mathcal{F}, e_{x,\infty})$ for $\epsilon > 0$. Then we define, for $n = 1, 2, \ldots$ and $\epsilon > 0$, the quantity

$$H_{n,\infty}(\epsilon, \mathcal{F}) = \sup_{x\in X^n} \log N(\epsilon, \mathcal{F}, e_{x,\infty})$$

The following is the relevant part of Theorem 6 from [21].

THEOREM 4.9. *Let $\mathcal{F}$ be a uniformly bounded family of functions on (X, Σ) that is also image admissible Suslin. Then $\mathcal{F}$ is a strong uniform Glivenko-Cantelli class if and only if*

$$\lim_{n\to\infty} H_{n,\infty}(\epsilon, \mathcal{F})/n = 0$$

for every $\epsilon > 0$.

5. Distance Approximation Schemes.

We saw in § 2.3 that amongst a number of bisimulation distance-estimation schemes for finite MDPs, the most promising appeared to be a method based on sampling. Therefore, we propose to extend this to the case of continuous MDPs. One would sample all probability mass functions, replace each with an empirical distribution built from the resulting samples, and repeatedly apply the fixed-point bisimulation functional to the new MDP. Supposing for the moment that one can enumerate and sample from a compact metric space with full-fledged probability measures, the only problem in this procedure is the validity of replacing the original MDP with the sampled version. In other words, if we replace the probability measures in our MDP with empirical measures, is it still true that the bisimulation metric on the sampled MDP will converge to the true bisimulation metric as the number of samples increases?

Fortunately, with some minor modifications the answer is yes. In order to prove this, we will need to make use of a uniform Glivenko-Cantelli theorem, Theorem 4.9. Such theorems typically characterize uniform convergence of empiricals to means, and are ubiquitous throughout machine learning [2]. Let us first take a moment to consider what this means in the context of the Kantorovich distances. Suppose $Lip(h)$ is a uniform Glivenko-Cantelli class for pseudometric h. Then the uniform Glivenko-Cantelli property tells us that $T_K(h)(\mu, \mu_i)$ converges to zero $\mathbb{P}$-almost surely for every μ *and* this convergence is uniform over all μ. Ideally then, we would like at least one of $Lip(\rho^*)$ and $Lip(\rho^*, C_b(S))$ to be a uniform Glivenko-Cantelli class. The question as to which classes constitute uniform Glivenko-Cantelli classes and under what conditions is an important area of active research. Fortunately, we have the following:

LEMMA 5.1. *Let (S, d) be a compact metric space and let $Lip(\rho^*, C_b(S))$ be as in Definition 2.38. Then $Lip(\rho^*, C_b(S))$ is a uniform Glivenko-Cantelli class.*

Proof. We will essentially follow the proof of Proposition 12 in [21]: if we can show that $Lip(\rho^*, C_b(S))$ is image admissible Suslin and that $\lim_{n\to\infty} H_{n,\infty}(\epsilon, Lip(\rho^*, C_b(S)))/n = 0$ for every $\epsilon > 0$ then the result will follow by Theorem 4.9.

As we saw in the proof of Lemma 3.14, $Lip(\rho^*, C_b(S))$ is equicontinuous at all points. It is uniformly bounded by $\|\rho^*\|$ by definition, and it is easily seen to be closed with respect to the uniform norm metric. Therefore, by Theorem 4.1 it is compact space, and hence also a Polish space and a Suslin space.

Now we have both (S, d) and $Lip(\rho^*, C_b(S))$ equipped with the uniform norm metric are compact, and hence separable metric spaces. Equip each with its Borel σ-algebra and note that by separability, the Borel σ-algebra of their product is the product of their Borel σ-algebras. Let T be the identity mapping on $Lip(\rho^*, C_b(S))$. Define $X : S \times Lip(\rho^*, C_b(S)) \to \mathbb{R}$ to be the evaluation map, that is, $X(f, s) = f(s)$. Define $\Gamma : S \times Lip(\rho^*, C_b(S)) \to \mathbb{R}$ by $\Gamma(f, s) = X(T(f), s) = f(s)$. Then since $Lip(\rho^*, C_b(S))$ is equicontinuous it follows that Γ is jointly continuous on the product space, and hence product measurable. Therefore, $Lip(\rho^*, C_b(S))$ is image admissible Suslin.

Finally note, that for any $n = 1, 2, \ldots$ and $s \in S^n$, $e_{s,\infty}$ is bounded above by the uniform norm metric. Thus, $N(\epsilon, Lip(\rho^*, C_b(S)), e_{s,\infty}) \le N(\epsilon, Lip(\rho^*, C_b(S)), \|\cdot\|)$, the latter term being finite and independent of n for every $\epsilon > 0$ as $Lip(\rho^*, C_b(S))$ with the uniform norm is a compact metric space. So for every $\epsilon > 0$, $H_{n,\infty}(\epsilon, Lip(\rho^*, C_b(S)))$ is uniformly bounded in n. Therefore, $\lim_{n\to\infty} H_{n,\infty}(\epsilon, Lip(\rho^*, C_b(S)))/n = 0$ for every $\epsilon > 0$, as was to be shown. □

How does this help us? Recall that as a first step in our distance approximation scheme, we would like to replace each probability measure on the space with an empirical measure and use Theorem 3.12 to guarantee existence of bisimulation metrics. However, in order to use that we require the map taking states to empirical measures to be continuous - and in general this need not be the case. We may circumvent this issue by replacing the Kantorovich operator with one that is defined on *all* real-valued functions, not just the measurable ones. For a fixed i, define for empiricals $\mu_i = \frac{1}{i}\sum_{j=1}^{i} \delta_{X_j}$ and $\nu_i = \frac{1}{i}\sum_{j=1}^{i} \delta_{Y_j}$ and bounded-pseudometric h,

$$T_K^i(h)(\mu_i, \nu_i) = \min_{\sigma} \frac{1}{i} \sum_{k=1}^{i} h(X_k, Y_{\sigma(k)})$$

(note that if h is measurable, then $T_K^i(h)(\mu_i, \nu_i) = T_K(h)(\mu_i, \nu_i)$). With this in mind, we may once more apply the Banach Fixed Point Theorem to obtain:

PROPOSITION 5.2. *Let $M = (S, \Sigma, A, P, r)$ be an MDP satisfying the conditions in Defini-*

tion 3.1, $c \in (0,1)$ be a metric discount factor, and $i \in \mathbb{N}$. Define $F_i : \mathfrak{met} \to \mathfrak{met}$ by

$$F_i(h)(s,s') = \max_{a \in A}((1-c)|r^a_s - r^a_{s'}| + cT^i_K(h)(P^a_{i,s}, P^a_{i,s'}))$$

Then :

1. *F_i has a unique fixed point ρ^*_i, and*
2. *for any $h_0 \in \mathfrak{met}$, $\lim_{n\to\infty}(F_i)^n(h_0) = \rho^*_i$.*

Note that technically, we have a random mapping here; that is, for each ω in Ω there is a mapping $F_i(\omega)$ from $\mathfrak{met}$ to itself with fixed point $\rho^*_i(\omega)$. So each ρ^*_i is really a (not necessarily measurable) mapping from Ω to $\mathfrak{met}$. Therefore, when speaking of convergence of the family $\{\rho^*_i\}_{i\in\mathbb{N}}$, we assume that convergence to be almost surely or in probability with respect to $\mathbb{P}^*$. We will omit the explicit use of ω in the rest of this work for the sake of convenience; however, the reader should make careful note of its existence.

Thus, the proposed statistical estimates $(\rho^*_i)_{i\in\mathbb{N}}$ to ρ^* exist; yet, how do we know that they actually converge to ρ^*? It is not hard to see that

$$\|\rho^*_i - \rho^*\| \leq \frac{2c}{1-c} \sup_{a\in A, s\in S} T_K(\rho^*)(P^a_{i,s}, P^a_s). \tag{5.1}$$

Simply note that

$$\begin{aligned}
|\rho^*_i(s,s') - \rho^*(s,s')| &\leq c \max_{a\in A} |T^i_K(\rho^*_i)(P^a_{i,s}, P^a_{i,s'}) - T_K(\rho^*)(P^a_s, P^a_{s'})| \\
&\leq c \max_{a\in A} |T^i_K(\rho^*_i)(P^a_{i,s}, P^a_{i,s'}) - T^i_K(\rho^*)(P^a_{i,s}, P^a_{i,s'})| \\
&\quad + c \max_{a\in A} |T_K(\rho^*)(P^a_{i,s}, P^a_{i,s'}) - T_K(\rho^*)(P^a_s, P^a_{s'})| \\
&\leq c\|\rho^*_i - \rho^*\| + c \max_{a\in A}(T_K(\rho^*)(P^a_{i,s}, P^a_s) + T_K(\rho^*)(P^a_{i,s'}, P^a_{s'})) \\
&\leq c\|\rho^*_i - \rho^*\| + 2c \sup_{a,s} T_K(\rho^*)(P^a_{i,s}, P^a_s)
\end{aligned}$$

and the result follows. This is where the uniform Glivenko-Cantelli property comes into play: we would like to use it to show that the quantity on the right-hand side of inequality 5.1 tends to zero almost surely. Unfortunately, we face a problem in the form of the supremum over the possibly uncountably infinite set S. While the uniform Glivenko-Cantelli theorem indeed tells us that empiricals converge in Kantorovich distance to their measure almost surely for each measure, and even over all measures almost surely for a countable set of measures, it does not dictate that all measures converge *at the same rate uniformly* almost surely. Here compactness comes to the rescue.

Let U be a countable dense subset of S, and let d be the metric on S. Recall that ρ^* is continuous on $S \times S$; in fact, since S is compact we may take ρ^* to be uniformly continuous. So for a fixed $\epsilon > 0$, there is a $\delta_c(\epsilon) > 0$ such that for any x, y in S, if $d(x,y) < \delta_c(\epsilon)$ then $\rho^*(x,y) < \epsilon$. In particular, we have

$$\max_{a\in A} T_K(\rho^*)(P^a_x, P^a_y) \leq \frac{1}{c}\rho^*(x,y) < \frac{\epsilon}{c}. \tag{5.2}$$

Let $[-] : S \to U$ be a mapping such that $d(s,[s]) < \delta_c(\epsilon)$ for every s in S and the image $[S]$ is finite; that this can be done is a consequence of U being dense in S and S being compact. Next, if

$\mu_i = \frac{1}{i}\sum_{j=1}^{i} \delta_{X_j}$, define $[\mu_i]$ to be $\frac{1}{i}\sum_{j=1}^{i} \delta_{[X_j]}$. Then for any μ_i

$$T_K(\rho^*)(\mu_i, [\mu_i]) = \min_{\sigma} \frac{1}{i}\sum_{k=1}^{i} \rho^*(X_k, [X_{\sigma(k)}]) \le \frac{1}{i}\sum_{k=1}^{i} \rho^*(X_k, [X_k]) < \epsilon \tag{5.3}$$

Now we are ready to proceed. The idea is that we will use statistical estimates of the probability measures as before; however, this time we will use $[-]$ to shift S to close-by points in U, thus restricting our calculations to the finite set $[S]$.

Theorem 5.3. *Let $M = (S, \Sigma, A, P, r)$ be an MDP satisfying the conditions in Definition 3.1, $c \in (0,1)$ be a metric discount factor, $i \in \mathbb{N}$, and $\epsilon > 0$. Further suppose that S is a compact metric space. Define $F_{i,\epsilon} : \mathfrak{met} \to \mathfrak{met}$ by*

$$F_{i,\epsilon}(h)(s, s') = \max_{a \in A}((1-c)|r^a_{[s]} - r^a_{[s']}| + cT^i_K(h)([P^a_{i,[s]}], [P^a_{i,[s']}]))$$

Then :

1. *$F_{i,\epsilon}$ has a unique fixed point $\rho^*_{i,\epsilon}$,*
2. *for any $h_0 \in \mathfrak{met}$, $\lim_{n\to\infty}(F_{i,\epsilon})^n(h_0) = \rho^*_{i,\epsilon}$, and*
3. *$\rho^*_{i,\epsilon}$ converges to ρ^* as $i \to \infty$ and $\epsilon \to 0$, $\mathbb{P}$-almost surely.*

Proof. The first two items once more follow from the Banach Fixed Point Theorem. As for the last item, let us show that

$$\|\rho^*_{i,\epsilon} - \rho^*\| \le \frac{1}{1-c}(2\epsilon(2+c) + 2c \max_{a\in A, u \in [S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)). \tag{5.4}$$

As in the previous proposition, let us note that

$$\begin{aligned}
|\rho^*_{i,\epsilon}(s,s') - \rho^*(s,s')| &\le (1-c)\max_{a\in A}(|r^a_{[s]} - r^a_{[s']}| - |r^a_s - r^a_{s'}|) \\
&\quad + c\max_{a\in A}|T^i_K(\rho^*_{i,\epsilon})([P^a_{i,[s]}], [P^a_{i,[s']}]) - T_K(\rho^*)(P^a_s, P^a_{s'})| \\
&\le \max_{a\in A}(1-c)|r^a_{[s]} - r^a_s| + \max_{a\in A}(1-c)|r^a_{[s']} - r^a_{s'}| \\
&\quad + c\max_{a\in A}|T^i_K(\rho^*_{i,\epsilon})([P^a_{i,[s]}], [P^a_{i,[s']}]) - T^i_K(\rho^*)([P^a_{i,[s]}], [P^a_{i,[s']}])| \\
&\quad + c\max_{a\in A}|T^i_K(\rho^*)([P^a_{i,[s]}], [P^a_{i,[s']}]) - T_K(\rho^*)(P^a_s, P^a_{s'})| \\
&\le \rho^*(s,[s]) + \rho^*(s',[s']) \\
&\quad + c\|\rho^*_{i,\epsilon} - \rho^*\| \\
&\quad + c\max_{a\in A}|T_K(\rho^*)([P^a_{i,[s]}], [P^a_{i,[s']}]) - T_K(\rho^*)(P^a_s, P^a_{s'})| \\
&\le \rho^*(s,[s]) + \rho^*(s',[s']) + c\|\rho^*_{i,\epsilon} - \rho^*\| \\
&\quad + c\max_{a\in A}\{T_K(\rho^*)([P^a_{i,[s]}], P^a_{i,[s]}) + T_K(\rho^*)(P^a_{i,[s]}, P^a_{[s]}) + T_K(\rho^*)(P^a_{[s]}, P^a_s) \\
&\quad + T_K(\rho^*)(P^a_{s'}, P^a_{[s']}) + T_K(\rho^*)(P^a_{[s']}, P^a_{i,[s']}) + T_K(\rho^*)(P^a_{i,[s']}, [P^a_{i,[s']}])\} \\
&\le \epsilon + \epsilon + c\|\rho^*_{i,\epsilon} - \rho^*\| \\
&\quad + c\max_{a\in A}\{\epsilon + T_K(\rho^*)(P^a_{i,[s]}, P^a_{[s]}) + \frac{\epsilon}{c} + \frac{\epsilon}{c} + T_K(\rho^*)(P^a_{[s']}, P^a_{i,[s']}) + \epsilon\} \\
&\le c\|\rho^*_{i,\epsilon} - \rho^*\| + 4\epsilon + 2c\epsilon + 2c\max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)
\end{aligned}$$

and the bound follows immediately. Note that in the fourth inequality we used the fact that ρ^* is meaurable to replace T_K^i by T_K, and in the fifth inequality we have repeatedly made use of inequalities 5.2 and 5.3.

By the Uniform Glivenko-Cantelli property, the rightmost term in inequality 5.4 tends to zero $\mathbb{P}$-almost surely (incidentally, dependent on ϵ); for, given a finite set $\mathcal{U}$ of measures, we have for a given $\varepsilon > 0$

$$\begin{aligned}&\mathbb{P}^*(\sup_{m\geq i}\sup_{\mu\in\mathcal{U}} T_K(\rho^*)(\mu_m,\mu) > \varepsilon) = \mathbb{P}^*(\sup_{\mu\in\mathcal{U}}\sup_{m\geq i} T_K(\rho^*)(\mu_m,\mu) > \varepsilon)\\ &\leq \sum_{\mu\in\mathcal{U}} \mathbb{P}^*(\sup_{m\geq i} T_K(\rho^*)(\mu_m,\mu) > \varepsilon)\\ &\leq |\mathcal{U}| \sup_{\mu} \mathbb{P}^*(\sup_{m\geq i} T_K(\rho^*)(\mu_m,\mu) > \varepsilon)\end{aligned}$$

whence it follows that $\mathbb{P}^*(\limsup_m \sup_{\mu\in\mathcal{U}} T_K(\rho^*)(\mu_m,\mu) > \varepsilon) = 0$. Since ε is arbitrary, we then have $\mathbb{P}^*(\limsup_m \sup_{\mu\in\mathcal{U}} T_K(\rho^*)(\mu_m,\mu) \neq 0) = 0$. Hence, for every $\epsilon > 0$

$$\lim_{i\to\infty} \|\rho^*_{i,\epsilon} - \rho^*\| \leq \frac{2\epsilon(2+c)}{1-c} \tag{5.5}$$

except for a set N_ϵ of $\mathbb{P}$-measure zero. Consider now only rational $\epsilon > 0$, and let N be the union of the collection $\{N_\epsilon\}$ over all such ϵ. Then save for N, inequality 5.5 holds for every ϵ, and N has $\mathbb{P}$-measure zero. So letting ϵ tend to zero in the same inequality, we find that $\rho_{i,\epsilon}$ converges to ρ^*, as was to be shown. □

Let us note that this then is the crucial result: it tells us that we may approximate ρ^* through $(F^n_{i,\epsilon})(\perp)$, that is, through sampling, discretization, and finite iteration and that we need only compute this latter quantity on $[S]$. More to the point, we may choose $[S] \subseteq U$ to be finite, since the $\delta(\epsilon)$-balls of U form an open cover of compact S. We now have the seeds of an algorithm.

6. Towards an Algorithm: Representation and Choice. We will assume we are provided with an "effective" representation of the state space S in terms of an enumeration of a countable dense subset U of S; we will additionally require that a specific metric d on S be specified as part of the input as a computable function on $U \times U$. The set of actions is simply a finite set A, and the reward function will be represented as an A-indexed family of computable functions from U to $[0,1]$. All that remains is to specify the transition probabilities.

How does one represent a probability measure on a continuous space? In the discrete setting, one of two approaches traditionally suffices: either probabilities can be specified point-to-point in a probability matrix, or one restricts attention to a parameterized class of probability mass functions. This latter approach also applies to Euclidean spaces, where one typically works with probability density functions. Although one may argue that both approaches can be suitably extended in the setting of a compact metric space (the interested reader is directed to the works of [22] and [62]) we will focus on the approach taken by [6].

Let us suppose that (S,d) is supplied with a canonical probability measure μ. We may then represent transition kernels inducing non-atomic probability measures by an A-indexed family of product measurable probability density functions, $f_a : S\times S \to [0,\infty)$, such $P^a_s(M) = \int_M f_a(s,\cdot)d\mu$. We will further suppose that $\mu(U) = 1$ and for each a, f_a is continuous in the first coordinate, and bounded by a μ-integrable function in the second; it then follows from the dominated convergence theorem that P^a_s is (weakly) continuous in s, and finally, that we need only specify each f_a on $U \times U$.

To summarize:

Definition 6.1. *A given continuous Markov decision process (S, Σ, A, P, r) with compact metric space (S, d) will be represented by the sextuple (U, d, μ, A, P, r), where:*

- *U is an enumeration of a countable dense subset of S,*
- *the metric d is computable on $U \times U$,*
- *μ is a canonical sampling measure on S satisfying $\mu(U) = 1$, and*
- *P_s^a is represented by*
  - ◇ *an atomic measure, given by a finite sum of point masses subject to the continuity constraint, or*
  - ◇ *a non-atomic measure, given by a probability density function $f_a : U \times U \to [0, \infty)$ continuous in the first coordinate, and bounded uniformly by a μ-integrable function in the second coordinate*
- *r is a computable function from $U \times A$ to $[0, 1]$ and continuous on U*

Lastly, we will assume that for a fixed positive rational ϵ we can enumerate a finite database $X \subseteq U$, such that the ϵ-balls centered at the points of X cover the entire space. Such an X is called an *ϵ-covering.* If X instead satisfies that all of its points are at least ϵ apart, then it is called an *$\frac{\epsilon}{2}$-packing.* The ideal situation would be one in which we can find an X that satisfies both properties; such an X is called an *ϵ-net.*

If a means of enumerating an ϵ-net for a given problem does not make itself obvious, then, as noted by [10], an ϵ-net X can be constructed by the following greedy algorithm essentially devised by [33] as an approximation algorithm for finding the smallest ϵ such that there is an ϵ-covering with k members, for a given k: given input $\epsilon > 0$ and maximum allowable ϵ-net size k, pick $s \in U$ arbitrarily, and set $X := \{s\}$. Then repeat the following: pick an $s \in U - X$ that maximizes $d(s, X) = \min\{d(s, x) : x \in X\}$. If $d(s, X) < \epsilon$ or $|X| \geq k$ then stop; otherwise, set $X := X \cup \{s\}$, and continue. Then X is an ϵ'-net for some $\epsilon' \leq \epsilon$ *provided* k is large enough; specifically, $\epsilon' := d(s, X - \{s\})$ where s was the last state to be added to X. If U is finite with size $n = |U|$ then this approximation algorithm has worst case running time $O(kn)$ with ϵ' within two times the optimal value.

The only problem in immediately applying this algorithm to the general case of a countably infinite U, is in finding the element s in U that maximizes $d(s, X)$. We can get around this by sampling according to μ: as suggested in [6] replace the maximum with the essential-supremum with respect to μ and approximate this via sampling according to μ and maximizing over the samples, which converges to the essential-supremum in probability. The resulting heuristic should provide an ample covering of U with high probability in time $O(kI)$ where I is the maximum number of samples used in estimating the essential-suprema provided an adequate number of samples is used. We provide no bound on the number of samples required here; we only note that there are methods for estimating an ϵ-net.

Our algorithm then is as follows: given a positive rational ϵ, enumerate a $\delta_c(\epsilon)$ cover X. Define $[s]$ to be the nearest neighbor of s in X according to d. Sample the probability distributions induced by X and use $[-]$ to restrict them to X. Finally, perform the iteration algorithm on X, as in the finite case. Figure 6.1 provides pseudocode for estimating distances to within an iteration error of δ for a given ϵ and ϵ-net X.

Theorem 6.2. *As in Definition 6.1, let (U, d, μ, A, P, r) be the respresentation of a given continuous MDP (S, Σ, A, P, r), where P is represented by the family of density functions $\{f_a : U \times U \to [0, \infty)\}_{a \in A}$. Let $c \in (0, 1)$ be a metric discount factor, $\epsilon > 0$ be a discretization parameter, δ be an iteration error, and i be the number of samples to be used in sampling P. Let $X \subseteq U$ be a*

```
INPUT: finite database $X \subseteq U$, finite action set $A$, number of samples $i$,
       reward function $r : U \times A \to [0,1]$, distance function $d : U \times U \to [0,\infty)$,
       density functions $\{f_a : U \times U \to [0,\infty)\}_{a \in A}$, sampling measure $\mu$,
       iteration error $\delta$

OUTPUT: distance function $\rho : X \times X \to [0,1]$

METHODS:
       NN$(z, d, X)$ returns nearest neighbor of $z$ in $X$ according to $d$.
       SAMPLE$(\mu, f)$ returns element of $U$ sampled independently according to
           probability measure induced by $\mu$ and density $f$.
       HUNGARIAN_ALG$(\rho, \vec{x}, \vec{y})$ returns value of minimum-cost assignment for
           assignment problem with cost $\rho$ and $i$-vectors $\vec{x}$ and $\vec{y}$ from $X$.

ALGORITHM:
(INITIALIZATION)
For $s, s' = 1$ to $|X|$ do
    $\rho(s, s') \leftarrow 0$
    For $a = 1$ to $|A|$ do
        For $j = 1$ to $i$ do
            $z \leftarrow$SAMPLE$(\mu, f_a(s, \cdot))$
            $P_a(s, j) \leftarrow$NN$(z, X, d)$
(MAIN LOOP)
For $j = 1$ to $\lceil \frac{\ln \delta}{\ln c} \rceil$ do
    For $s, s' = 1$ to $|X|$ do
        For $a = 1$ to $|A|$ do
            $TK_a(s, s') \leftarrow$HUNGARIAN_ALG$(\rho, P_a(s, \cdot), P_a(s', \cdot))$
    For $s, s' = 1$ to $|X|$ do
        $\rho(s, s') \leftarrow \max_a((1-c)|r(s,a) - r(s',a)| + cTK_a(s, s'))$
```

FIG. 6.1. *Pseudocode for estimating bisimulation distances*

finite database that is an ϵ-cover of S. Then the algorithm given by the pseudocode in Figure 6.1 computes an approximation $\rho : X \times X \to [0,1]$ to the bisimulation metric ρ^* given by Theorem 3.12 in worst case running time $O(\frac{\ln \delta}{\ln c} mn^2 i^3)$ and with error bounded above by

$$\delta + \frac{2\epsilon(2+c)}{1-c} + \frac{2c}{1-c} \max_{a \in A, x \in X} T_K(\rho^*)(P^a_{i,x}, P^a_x)$$

The next section is dedicated towards verifying the bounds on the running time and the approximation error, and trying to further provide error estimation guarantees.

7. Estimation Error. Let us analyze the error of our approximation algorithm for the 1-bounded bisimulation metric ρ^*. Recall that we are approximating ρ^* by $F^n_{i,\epsilon}(\perp)$ for large i and n, and small ϵ. So the approximation error is given by:

$$\begin{aligned}\|(F_{i,\epsilon})^n(\bot) - \rho^*\| \le \|(F_{i,\epsilon})^n(\bot) - \rho^*_{i,\epsilon}\| + \|\rho^*_{i,\epsilon} - \rho^*\| &\le \frac{c^n}{1-c}\|F_{i,\epsilon}(\bot)\| + \|\rho^*_{i,\epsilon} - \rho^*\| \\ &\le \frac{c^n}{1-c}(1-c) + \frac{1}{1-c}(2\epsilon(2+c) + 2c \max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)) \\ &\le c^n + \frac{2\epsilon(2+c)}{1-c} + \frac{2c}{1-c} \max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)\end{aligned}$$

Let $\varepsilon_\sim$, $\varepsilon_{[-]}$, and $\varepsilon_{\mathbb{P}}$ denote c^n, $\frac{2\epsilon(2+c)}{1-c}$, and $\frac{2c}{1-c}\max_{a\in A,u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)$; these are, respectively, the bisimilarity, discretization, and sampling errors. In the next few sections, we will try to bound these to within some prescribed degree of accuracy.

7.1. Bisimulation Error. Bounding the error due to approximating bisimilarity in n steps is simple enough. Suppose we want this error to be less than δ for some $\delta > 0$. Choose $n = \lceil \frac{\ln \delta}{\ln c} \rceil$; then $\varepsilon_\sim = c^n \le c^{\frac{\ln \delta}{\ln c}} = e^{\ln \delta} = \delta$. So we need only iterate for $\lceil \frac{\ln \delta}{\ln c} \rceil$ steps.

7.2. Discretization Error. In some sense, bounding the discretization error is hopeless - we need to know how ρ^* varies with d and in general, this is information that we just do not have. However, there is some hope; recall that what we need is some way of specifying a $\delta_c(\epsilon)$ so that $d(x,y) < \delta_c(\epsilon)$ implies $\rho^*(x,y) < \epsilon$. Suppose we can bound ρ^* from above by a continuous metric m; define the modified metric d_m to be $\max(d,m)$. Then, as $d \le d_m$ and d_m is continuous with respect to d, we have that d_m and d are equivalent metrics; that is, they induce the same topology on S. Therefore, we could use d_m in place of d and simply take $\delta_c(\epsilon)$ to be ϵ; but how do we find m? More to the point - as ρ^* is itself a candidate - how do we find an m that is easier to compute than ρ^*?

We propose here a heuristic for computing such an m. We cannot hope to bound the discretization error in computing m owing to the reasons mentioned above; however, we hope to shift the focus of the discretization error onto how r and P vary with d. In other words, if we discretize the state space using an ϵ-net with respect to d_m then we will be able to set $\varepsilon_{[-]} = \frac{2\epsilon(2+c)}{1-c} + \varepsilon_m$ where ε_m, the estimation error for d_m, hopefully varies much more closely with d than does ρ^*.

Let $\mathfrak{cts_m} \subseteq \mathfrak{lsc_m}$ denote the space of bounded continuous pseudometrics on S. Define $R \in \mathfrak{cts_m}$ and the operator $T : \mathfrak{cts_m} \to \mathfrak{cts_m}$ by

$$R(x,y) = \max_{a\in A} |r^a_x - r^a_y| \text{ and } T(h)(x,y) = \max_{a\in A}(P^a_x \otimes P^a_y)(h),$$

where $\mu \otimes \nu$ is the product measure of μ and ν. The fact that $T(h)$ is symmetric follows from the Fubini-Tonelli Theorem (see for example [29]), which allows one to change the order of integration in an iterated integral. The fact that $T(h)$ is continuous for h in $\mathfrak{cts_m}$ follows from the fact that for separable metric spaces the limit of the product of weakly converging measures is the product of the limits of those measures: if $\mu_n \Rightarrow \mu$ and $\nu_n \Rightarrow \nu$ then $\mu_n \otimes \nu_n \Rightarrow \mu \otimes \nu$ [4]. We immediately have that for any $h \in \mathfrak{cts_m}$, $F(h) \le (1-c)R + cT(h)$, where F is the fixed point operator for ρ^*. Finally, we define

$$m := (1-c)\sum_{k=0}^{\infty} c^k T^k(R).$$

Note that by comparison with the geometric series $(1-c)\sum_{k=0}^{\infty} c^k$, m converges absolutely everywhere. Moreover, as the sequence of partial sums belong to $\mathfrak{cts}_{\mathfrak{m}}$ and converge uniformly to m, m too belongs to $\mathfrak{cts}_{\mathfrak{m}}$. Now for any x,y and a, the Monotone Convergence Theorem tells us that

$$(P_x^a \otimes P_y^a)(m) = (P_x^a \otimes P_y^a)((1-c)\sum_{k=0}^{\infty} c^k T^k(R)) = (1-c)\sum_{k=0}^{\infty} c^k (P_x^a \otimes P_y^a)(T^k(R)).$$

Hence, taking the maximum over all actions yields $T(m) \leq (1-c)\sum_{k=0}^{\infty} c^k T^{k+1}(R)$. Thus, $F(m) \leq (1-c)R + cT(m) \leq (1-c)R + c(1-c)\sum_{k=0}^{\infty} c^k T^{k+1}(R) = m$, whence it follows that $\rho^* \leq m$.

Let us assume that we can compute $(P_x^a \otimes P_y^a)(h)$ for any computable h, for example, through numerical integration, sampling, etc. Then we can compute m for any pair of states by iterating T until c^n is less than some prescribed degree of accuracy and computing the nth partial sum. This, of course, introduces the additional estimation error ε_m. Finally, d_m can be computed as the maximum of m and d, and can even be taken to be 1-bounded since m is bounded by 1. For example, we may replace d with the compatible 1-bounded metrics $\frac{d}{1+d}$ or $\min(1,d)$.

7.3. Sampling Error. Let us first note that, strictly speaking, the expression denoted by $\varepsilon_{\mathbb{P}}$ is not solely the error due to sampling; for it is dependent on the measures indexed by $[S]$, i.e. it measures error due to discretization as well. In addition, though this term does tend to zero almost surely, it will be easier in practice to bound its convergence in probability. Let us suppose we want $\varepsilon_{\mathbb{P}}$ to be less than or equal to Δ with probability at least $1-\alpha$ for some small positive constants Δ and α. Note that

$$\begin{aligned}\mathbb{P}^*(\varepsilon_{\mathbb{P}} > \Delta) &= \mathbb{P}^*\big(\max_{a\in A, u\in[S]} T_K(\rho^*)(P_{i,u}^a, P_u^a) > \frac{1-c}{2c}\Delta\big)\\ &\leq |A||[S]| \sup_{a\in A, u\in[S]} \mathbb{P}^*\big(T_K(\rho^*)(P_{i,u}^a, P_u^a) > \frac{1-c}{2c}\Delta\big).\end{aligned}$$

Thus, it will suffice to find a uniform Glivenko-Cantelli convergence bound for

$$\sup_{u\in[S]} \mathbb{P}^*\big(T_K(\rho^*)(P_{i,u}^a, P_u^a) > \frac{1-c}{2c}\Delta\big) \leq \frac{\alpha}{|A||[S]|}. \tag{7.1}$$

The lower bound on the number of samples required to achieve the specified level of accuracy with the specified probability is known as the *sample complexity*. A large number of bounds exist in terms of various notions of dimension: VC-dimension, the fat-shattering dimension, covering numbers [2]; in general, a specific bound will depend on the structure of the metric space in question. As such, we are not able to provide specific bounds for the sample complexity in full generality. However, as an example, the following asymptotic lower bound for 7.1 can be obtained from Theorem 3.6 of [1]:

$$i = O\big(\frac{1}{\varepsilon^2}\big(\beta \ln^2 \frac{\beta}{\varepsilon} + \ln\frac{1}{\eta}\big)\big),$$

where $\varepsilon = \frac{1-c}{2c}\Delta$, $\eta = \frac{\alpha}{|A||[S]|}$, and β is the fat-shattering dimension of $Lip(\rho^*, C_b(S))$ with scale $\frac{\epsilon}{24}$: for a given class $\mathfrak{F}$ of $[0,1]$-valued functions on S and a given positive real number γ, one says $S' \subseteq S$ is γ-shattered by $\mathfrak{F}$ if there exists a function $s : S' \to [0,1]$ such that for every $S'' \subseteq S'$ there exists some $f_{S''} \in \mathfrak{F}$ that satisfies for every $x \in S'\backslash S''$, $f_{S''}(x) \leq s(x) - \gamma$ and for every $x \in S''$, $f_{S''}(x) \geq s(x) + \gamma$. The fat-shattering dimension of $\mathfrak{F}$ at scale γ is the maximum cardinality of a γ-shattered set.

7.4. Computational Complexity. Precise computational complexity results are difficult to come by owing to the application of this work to general metric spaces. The particular performance will depend on the structure of a given space – and this in turn can be represented by a number of proposed measures of metric space dimension [10]. However, the previous sections do give an idea of the space and time requirements in computing distances to a given level of accuracy with a given probability. A quick glance will tell us that it would be very expensive to attempt to compute distances to within a very small degree of error with high probability – but this is none too surprising. Previous work [59] has shown that computing the bisimulation distances for a given finite probabilistic system in tabular form can be done in polynomial time. In practice we fix the number of samples in our sampling procedure and sacrifice accuracy for improved running times; that is, for a fixed number of samples i and a given discretization $[-]$, let n be the number of discretized states in $[S]$ and m be the number of actions; then computing the state-similarity distances to within a bisimilarity error of δ requires time $O(\frac{\ln\delta}{\ln c}mn^2i^3)$. In order to see this, let us refer to the pseudocode in Figure 6.1: in the initialization phase, for every discrete state and for every action, a sample is obtained and a nearest neighbour search is peformed, i times. let us assume that sampling takes constant time; then this requires $O(nmi(O(1)+n))$, or $O(mn^2i)$ steps. In the algorithm's main loop, we iterate the following procedure for $\lceil\frac{\ln\delta}{\ln c}\rceil$ steps: for every pair of states and for every action, peform the Hungarian algorthim on their induced empirical probability distributions, taking $O(i^3)$ steps for each pair and leading to a total of $O(n^2mi^3)$ steps. Then for every pair of states a maximization must be performed over the m actions, requiring a total of $O(n^2m)$ steps. So the main loop requires $O(\frac{\ln\delta}{\ln c}(mn^2i^3+mn^2))$, or $O(\frac{\ln\delta}{\ln c}mn^2i^3)$ steps. The entire algorithm then requires $O(mn^2i)+O(\frac{\ln\delta}{\ln c}mn^2i^3) = O(\frac{\ln\delta}{\ln c}mn^2i^3)$ steps. Future algorithmic efficiency, however, will require the imposition of several structural/representational conditions and learning just how to exploit these.

8. Conclusions. In this paper we have established a robust quantitative analogue of bisimilarity for Markov decision processes with continuous state spaces in the form of a continuous pseudometric on the system states. More importantly, we have developed a novel distance-estimation scheme for MDPs with compact metric state spaces, permitting for what we believe is the first time the use of metric based reasoning for continuous probabilistic systems in practice.

The ability to estimate bisimulation distances for a wide class of continuous systems provides an invaluable tool for finding solutions to a similarly wide class of problems. One can compare the performance of several candidate state aggregation schemes in practice, or one can use the distances themselves to aggregate; in either case the distances provide meaningful error bounds on the quality of the models. Equally important, they provide tight error bounds on the quality of solutions obtained from finite approximations through the continuity bounds we've obtained on the optimal value function.

8.1. Related Work. This work has its roots in the work of Desharnais et al. [17] and van Breugel and Worrell [59]. In the work of [16, 17] and mainly in the thesis of [14], the authors developed bisimulation metrics for a probabilistic transition model similar to the Markov decision process, namely the labeled Markov process (LMP) [5]:

Definition 8.1. *A* labeled Markov process *is a quadruple*

$$(S, \Sigma, A, \{\tau_a | a \in A\})$$

where:

- *S is an* analytic *set of states*

- *Σ is the Borel σ-field on S*
- *A is a finite set of actions*
- *for every $a \in A$, $\tau_a : S \times \Sigma \to [0,1]$ is a stationary subprobability transition kernel:*
  - $\diamond$ *for every $X \in \Sigma$, $\tau_a(\cdot, X)$ is a measurable function and*
  - $\diamond$ *for every $s \in S$, $\tau_a(s, \cdot)$ is a subprobability measure*

An LMP can best be thought of here as a continuous state space MDP, with the difference being that it allows for subprobability measures and lacks rewards. It is worth noting that the authors develop their theory in the slightly more general setting of analytic spaces.

One may define bisimilarity for an LMP as follows: Recall Definition 3.2. A *bisimulation relation* is an equivalence relation R on S that satisfies the following property:

$$sRs' \iff \text{ for every } a \in A \text{ and } R\text{-closed } X \in \Sigma, \tau_a(s,X) = \tau_a(s',X)$$

We say states two states are *bisimilar* if and only if they are related by some bisimulation relation.

One may also define bisimilarity for LMPs in terms of a modal logic: two states are bisimilar if and only if they satisfy exactly the same formulas in some fixed logic [5, 14]. This forms the basis for the metrics of [14, 16, 17], which are defined in terms of real-valued logical expressions. The intuition in moving to metrics is that the bisimilarity of two states is directly related to the complexity of the simplest formula that can distinguish them; the "more bisimilar" two states are, the harder it should be to find a distinguishing formula; hence, such a formula should be necessarily "big". Of course, to formalize this one needs to find some quantitative analogue of logical formulas and satisfaction. One idea of how to do this in the context of a probabilistic framework comes from [39]:

Classical Logic	Generalization
Truth values 0,1	Interval [0,1]
Propositional function	Measurable function
State	Measure
The satisfaction relation $\models$	Integration $\int$

The idea is that just as the satisfaction relation maps states and propositional formulas to truth values, integration maps measures and measurable functions to extended truth values - values in the closed unit interval $[0,1]$. On the basis of these ideas, [14] developed a class of logical functional expressions that could be evaluated on the state space of a given LMP to obtain values in the unit interval. A family of bisimulation metrics is then constructed by calculating the difference of these quantities for a fixed pair of states across all formulas. Formally, let $c \in (0,1]$ and let $\mathcal{F}^c$ be a family of functional expressions whose syntax is given by the following grammar:

$$f := \mathit{1} | \min(f,f) | \langle a \rangle f | f \ominus q | \lceil f \rceil^q$$

where a and q range over A and rationals in $[0,1]$ respectively. These functional expressions are evaluated on S as follows:

$$\begin{aligned}
\mathit{1}(s) &= 1\\
\min(f_1, f_2)(s) &= \min(f_1(s), f_2(s))\\
(\langle a\rangle f)(s) &= c\int_S f(x)\tau_a(s, dx)\\
(f \ominus q)(s) &= \max(f(s) - q, 0)\\
\lceil f\rceil^q(s) &= \min(f(s), q)
\end{aligned}$$

Lastly, define $d^c : S \times S \to [0, 1]$ by

$$d^c(s, s') = \sup_{f\in\mathcal{F}^c} |f(s) - f(s')|.$$

THEOREM 8.2 ([14]). *For every c in $(0, 1]$, d^c is a 1-bounded bisimulation metric.*

In the finite case and with $c < 1$, Desharnais et al. [16] were able to construct a decision procedure for computing the metrics to any desired accuracy; one simply replaces $\mathcal{F}^c$ in the definition above with a specially chosen finite subset of functions. However, in the general case no algorithm was provided and it remained unclear as to whether or not d^1 was computable.

Later on, van Breugel and Worrell [58, 59] worked with a slightly modified version of these metrics in a categorical setting; they used fixed point theory in conjunction with the Kantorovich probably metric to define metrics on LMPs. They were able to show that the metrics induced by the logical characterization of bisimilarity and provided by Desharnais et al. [16] coincided with their own fixed point metrics. Particularly important was their application of the Kantorovich operator and subsequent use of network linear programming to develop the first polynomial-time decision procedure for the metrics in the finite case. In recent years, van Breugel, Sharma and Worrell [57] have developed both a theoretical framework and a decision procedure for finite LMP metrics without discounting, that is, for $c = 1$. Still, no work has been carried out on estimating distances for general LMPs with continuous state spaces.

In the context of MDPs, a number of methods have been proposed for analyzing state-similarity. Li, Walsh and Littman [42], for example, survey a number of state aggregation techniques for finite MDPs in an attempt to unify the theory of state abstraction: these include aggregation of states based on bisimulation, homomorphisms, value equivalence, and policy equivalence, to name a few. Muller [46] gave an early sensitivity analysis result in a spirit very similar to our own; he considers abstract MDPs (with full measurable state and action spaces) in which only the stochastic transition kernels differ. He then demonstrates continuity of a sort for the optimal value function with respect to several integral probability metrics. However, these results are purely of a mathematical nature – no algorithm is provided or even suggested. Goubault-Larrecq [35] introduces a hemi-metric (such a function satisfies all properties of a pseudometric save for symmetry) for simulation in *prevision transition systems*, a generalization of probabilistic transitions systems. There the setting is again continuous state spaces, and the author presents a similar value function continuity result (Proposition 4 of [35]) to Theorem 3.20 under continuity conditions similar to those found in Definition 3.1.

In the realm of finite MDPs, several works have analyzed the error in perturbing the parameters of a given Markov decision process. Dean, Givan and Leach [12] consider bounded-parameter MDPs, in which reward and probability parameters are specified by intervals of closed reals, and define ϵ-homogeneity: a loosening of bisimulation such that all states in the same equivalence class have reward parameters and probability parameters each differing by at most ϵ. In the paper of Even-Dar and Mansour [23], this work was expanded upon by considering different norms on the probability parameter in the definition of ϵ-homogeneity and providing performance results specifically showing that the quality of an ϵ-homogeneous partition depended heavily on the norm in use. More recently, Ortner [48] has expanded upon the notion of ϵ-homogeneity in terms of *adequate* pseudometrics and used these results to analyze finite MDPs under an average reward optimality criterion.

8.2. Future Work. There are many interesting directions possible for future investigation. Chief among these is the question of whether or not the results appearing in this work remain valid with less stringent or alternative conditions on the Markov decision problem parameters. Let us make a few quick remarks on this matter: firstly, the work of Desharnais et al. [14] for LMPs provides ample evidence that existence of our metrics should remain valid in at least analytic spaces. Following along the lines of Muller [46], we may replace uniform boundedness of rewards with boundedness in terms of a bounding "weight" function, which controls the rate at which the functions grow - this essentially amounts to replacing all uniform norms by weighted uniform norms in the proofs of this work. Promising work on Kantorovich duality [13] may allow us to show that the mapping of states to the Kantorovich distance of their induced distributions in Theorem 3.12 is a measurable mapping, thereby allowing us to remove continuity conditions on the reward and probability parameters, at least in existence proofs.

There are problem instances where each time step is equally important, and discounting is unsuitable; in these cases an average reward optimality criterion [53] is preferable for finding optimal polices for a given Markov decision process. We conjecture that $\lim_{c \to 1} \rho^*$ may yield a bisimulation metric suitable for analyzing average reward Markov decision problems.

We could also consider applying our work to extensions of bisimilarity. Desharnais et al. [18], for example, utilize *weak bisimulation* instead of bisimulation when developing a quantitative notion of state-similarity for a finite probabilistic transition system: essentially, states are deemed equivalent if they match over a sequence of transitions, rather than precisely at every step.

An immediate concern is that the algorithm proposed in this work was tested merely to illustrate its validity; a more extensive investigation will be carried out at a later stage. In practice, however, MDPs are rarely represented explicitly; instead, researchers usually work with factored representations [9], wherein the state space is represented by a family of state variables. Each MDP parameter is then compactly represented in terms of these variables, for example through use of dynamic Bayes nets or multi-terminal binary decision diagrams, yielding a compact representation of an MDP. If metric calculation can be adapted to work solely with the factored representation, and it is our strong belief that this is the case, then one would expect a great savings in the performance of such state-similarity algorithms.

Another natural extension is to apply this work to partially observable MDPs (POMDPs) [37]. A POMDP basically consists of an MDP in which the actual state of the system is hidden; instead one has a visible set of observations and a probabilistic observation function. A finite POMDP is a sextuple

$$(S, A, \{r_s^a | s \in S, a \in A\}, \{P_{ss'}^a | s, s' \in S, a \in A\}, \Omega, \{O_{so}^a | s, \in S, o \in O, a \in A\})$$

where:

- $(S, A, \{r_s^a | s \in S, a \in A\}, \{P_{ss'}^a | s, s' \in S, a \in A\})$ is a finite MDP
- Ω is a finite set of observations, and
- for every $s \in S$, $o \in \Omega$, and $a \in A$, O_{so}^a is the probability of observing observation o after a transition to state s under action a

Each POMDP induces a continuous state-space MDP from which a solution may be recovered. This continuous MDP, the *belief state MDP*, is given by

$$(\mathcal{B}, A, \{r^a_b | b \in \mathcal{B}, a \in A\}, \{P^a_{bb'} | b, b' \in \mathcal{B}, a \in A\})$$

where:

- $\mathcal{B}$ is the set of belief states on S, where a belief state b is defined to be a probability distribution on S,
- A is the same set of actions
- for each $b \in \mathcal{B}$, $a \in A$,

$$r^a_b = \sum_{s \in S} r^a_s b(s)$$

- for each $b, b' \in \mathcal{B}$, $a \in A$,

$$P^a_{bb'} = \sum_{o \in \Omega} Pr(b'|a, b, o) \sum_{s' \in S} O^a_{s'o} \sum_{s \in S} P^a_{ss'} b(s)$$

  where $Pr(b'|a, b, o) = 1$ if $b' = b_{(a,b,o)}$ and 0 otherwise, and

$$b_{(a,b,o)}(s') = \frac{O^a_{s'o} \sum_{s \in S} P^a_{ss'} b(s)}{Pr(o|a, b)}$$

  the denominator being calculated as a normalizing constant.

Optimal policies for the belief state MDP are optimal policies for the original POMDP. In this sense, our results for continuous MDPs would immediately apply; however, a more direct solution would be preferable.

The most evident use of our metrics is in analyzing state aggregations; however, the original motivation for a quantitative notion of bisimilarity was to study performance properties of a system, specified in terms of a modal logic [16, 14]. In fact, the original LMP metrics were defined in terms of a real-valued modal logic that captured properties of the system's states. Though we have not covered the logical approach for the continuous case in this work, it should easily be carried over with only slight modification. Thus, our metrics have a potential use in reasoning about logical properties of continuous MDPs too.

There has also been some preliminary work on knowledge transfer of policies in MDPs [52]. The basic idea is that if two MDPs have small overall bisimulation distance then how close a policy is to optimality in one model bounds how close it is to optimality in the other:

Theorem 8.3 ([52]). *Suppose $M_i = (S_i, A, \{r^a_{i,s} | s \in S, a \in A\}, \{P^a_{i,ss'} | s, s' \in S, a \in A\})$ are two finite MDPs for $i = 1, 2$ and suppose further there is a mapping $[-] : S_1 \to S_2$ specifying for each state in M_1 its representative state in M_2. Any policy π_2 defined on M_2 naturally defines a policy $\pi_1$on M_1 given by $\pi_1(s, a) = \pi_2([s], a)$, and in this way one can transfer policy π_2 from M_2 to M_1. Let γ and c be value and metric discount factors in $(0, 1)$ respectively, with $\gamma \leq c$. Let V^*_1 and V^*_2 be the optimal value functions for M_1 and M_2, respectively. Let ρ^* be the bisimulation metric defined on the disjoint union of M_1 and M_2. Then*

$$\|V^{\pi_1} - V^*_1\| \leq 2 \max_{s \in S_1} \rho^*(s, [s]) + \frac{1 + c}{1 - c} \|V^{\pi_2} - V^*_2\|$$

One could potentially solve a class of MDPs by using the solution to a base MDP to which they are all similar, and modifying that policy accordingly.

Finally, it is natural to consider two extensions: models with continuous time and continuous action spaces. This, in conjunction with the current work on continuous state spaces, is the subject of ongoing work.

REFERENCES

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler, *Scale-sensitive dimensions, uniform convergence, and learnability*, Journal of the ACM (JACM), 44 (1997), pp. 615–631.

[2] Martin Anthony, *Uniform Glivenko-Cantelli theorems and concentration of measure in the mathematical modelling of learning*, Tech. Report LSE-CDAM-2002-07, Centre for Discrete and Applicable Mathematics, 2002. Also found at: www.maths.lse.ac.uk/Personal/martin/mresearch.html.

[3] Marco Bernardo and Mario Bravetti, *Performance measure sensitive congruences for Markovian process algebras*, Theoretical Computer Science, 290 (2003), pp. 117–160.

[4] P. Billingsley, *Convergence of Probability Measures*, Wiley, 1968.

[5] Richard Blute, Josée Desharnais, Abbas Edalat, and Prakash Panangaden, *Bisimulation for labelled Markov processes*, in LICS '97: Proceedings of the 12th Annual IEEE Symposium on Logic in Computer Science, Washington, DC, USA, 1997, IEEE Computer Society, pp. 149–159.

[6] Alexandre Bouchard-Côté, Norm Ferns, Prakash Panangaden, and Doina Precup, *An approximation algorithm for labelled Markov processes: Towards realistic approximation*, in QEST '05: Proceedings of the Second International Conference on the Quantitative Evaluation of Systems (QEST'05) on The Quantitative Evaluation of Systems, Washington, DC, USA, 2005, IEEE Computer Society, pp. 54–61.

[7] Craig Boutilier, Thomas Dean, and Steve Hanks, *Decision-theoretic planning: Structural assumptions and computational leverage*, Journal of Artificial Intelligence Research, 11 (1999), pp. 1–94.

[8] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt, *Exploiting structure in policy construction*, in Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Chris Mellish, ed., San Francisco, 1995, Morgan Kaufmann, pp. 1104–1111.

[9] ———, *Stochastic dynamic programming with factored representations*, Artificial Intelligence, 121 (2000), pp. 49–107.

[10] Kenneth L. Clarkson, *Nearest-neighbor searching and metric space dimensions*, in Nearest-Neighbor Methods for Learning and Vision: Theory and Practice, Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk, eds., MIT Press, 2006, pp. 15–59.

[11] Thomas Dean and Robert Givan, *Model minimization in Markov decision processes*, in Association for the Advancement of Artificial Intelligence AAAI/ Innovative Applications of Artificial Intelligence IAAI, 1997, pp. 106–111.

[12] Thomas Dean, Robert Givan, and Sonia Leach, *Model reduction techniques for computing approximately optimal solutions for Markov decision processes*, in Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97), San Francisco, CA, 1997, Morgan Kaufmann, pp. 124–131.

[13] J. Dedecker, C. Prieur, and P. Raynaud De Fitte, *Parametrized Kantorovich-Rubinstein theorem and application to the coupling of random variables*, ArXiv Mathematics e-prints, (2004).

[14] Josée Desharnais, *Labelled Markov Processes*, PhD thesis, McGill University, 2000.

[15] J. Desharnais, A. Edalat, and P. Panangaden, *Bisimulation for labeled Markov processes*, Information and Computation, 179 (2002), pp. 163–193.

[16] Josée Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden, *Metrics for labeled Markov systems*, in CONCUR '99: Proceedings of the 10th International Conference on Concurrency Theory, London, UK, 1999, Springer-Verlag, pp. 258–273.

[17] ———, *Metrics for labelled Markov processes*, Theor. Comput. Sci., 318 (2004), pp. 323–354.

[18] Josée Desharnais, Radha Jagadeesan, Vineet Gupta, and Prakash Panangaden, *The metric analogue of weak bisimulation for probabilistic processes*, in LICS '02: Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science, Copenhagen, Denmark, 22-25 July 2002, Washington, DC, USA, 2002, IEEE Computer Society, pp. 413–422.

[19] R. M. Dudley, *Notes on empirical processes*, 2000. Lecture notes for a course given at Aarhus University, August 1999.

[20] ———, *Real Analysis and Probability*, Cambridge University Press, August 2002.

[21] R. M. Dudley, E. Giné;, and J. Zinn, *Uniform and universal Glivenko-Cantelli classes*, Journal of Theoretical

Probability, 4 (1991), pp. 485–510.

[22] Abbas Edalat, *When Scott is weak on the top*, Mathematical Structures in Computer Science, 7 (1997), pp. 401–417.

[23] Eyal Even-Dar and Yishay Mansour, *Approximate equivalence of Markov decision processes*, in 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings, vol. 2777 of Lecture Notes in Computer Science, Springer, 2003, pp. 581–594.

[24] Norm Ferns, *Metrics for Markov decision processes*, Master's thesis, McGill University, Montreal, Canada, 2003.

[25] ———, *State-Similarity Metrics for Continuous Markov Decision Processes*, PhD thesis, McGill University, Montreal, Canada, 2008.

[26] Norm Ferns, Pablo Samuel Castro, Doina Precup, and Prakash Panangaden, *Methods for computing state similarity in Markov decision processes*, in Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06), Arlington, Virginia, 2006, AUAI Press.

[27] Norm Ferns, Prakash Panangaden, and Doina Precup, *Metrics for finite Markov decision processes*, in AUAI '04: Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence, Arlington, Virginia, United States, 2004, AUAI Press, pp. 162–169.

[28] ———, *Metrics for Markov decision processes with infinite state spaces*, in Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05), Arlington, Virginia, 2005, AUAI Press, pp. 201–208.

[29] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Wiley-Interscience, second ed., 1999.

[30] Antonio Frangioni and Antonio Manca, *A computational study of cost reoptimization for min-cost flow problems*, INFORMS J. on Computing, 18 (2006), pp. 61–70.

[31] Alison L. Gibbs and Francis Edward Su, *On choosing and bounding probability metrics*, International Statistical Review, 70 (2002), pp. 419–435.

[32] Robert Givan, Thomas Dean, and Matthew Greig, *Equivalence notions and model minimization in Markov decision processes*, Artificial Intelligence, 147 (2003), pp. 163–223.

[33] Teofilo F. Gonzalez, *Clustering to minimize the maximum intercluster distance*, Theoretical Computer Science, 38 (1985), pp. 293–306.

[34] Jason H. Goto, Mark E. Lewis, and Martin L. Puterman, *Coffee, tea, or ...?: A Markov decision process model for airline meal provisioning*, Transportation Science, 38 (2004), pp. 107–118.

[35] Jean Goubault-Larrecq, *Simulation hemi-metrics between infinite-state stochastic games*, in Proceedings of the Theory and Practice of Software, 11th International Conference on Foundations of Software Science and Computational Structures, FOSSACS'08/ETAPS'08, Berlin, Heidelberg, 2008, Springer-Verlag, pp. 50–65.

[36] Matthew Hennessy and Robin Milner, *Algebraic laws for nondeterminism and concurrency*, Journal of the Association for Computing Machinery (JACM), 32 (1985), pp. 137–161.

[37] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra, *Planning and acting in partially observable stochastic domains*, Artificial Intelligence, 101 (1998), pp. 99–134.

[38] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Van Nostrand, 1960.

[39] Dexter Kozen, *A probabilistic PDL*, in STOC '83: Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing, New York, NY, USA, 1983, ACM, pp. 291–297.

[40] T. Lane and L. Pack Kaelbling, *Approaches to macro decompositions of large Markov decision process planning problems*, in Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference on Mobile Robots XVI, Howie. M. Gage, Douglas W.and Choset, ed., vol. 4573 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Newton, MA, feb 2002, SPIE, pp. 104–113.

[41] Kim G. Larsen and Arne Skou, *Bisimulation through probabilistic testing*, Information and Computation, 94 (1991), pp. 1–28.

[42] Lihong Li, Thomas J. Walsh, and Michael L. Littman, *Towards a unified theory of state abstraction for MDPs*, in AI & MATH '06: Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics, 2006, pp. 531–539.

[43] Maxim Likhachev, Geoff Gordon, and Sebastian Thrun, *Planning for Markov decision processes with sparse stochasticity*, in Advances in Neural Information Processing Systems 17, Lawrence K. Saul, Yair Weiss, and Léon Bottou, eds., MIT Press, Cambridge, MA, 2005, pp. 785–792.

[44] Robin Milner, *A Calculus of Communicating Systems*, vol. 92 of Lecture Notes in Computer Science, Springer-Verlag, New York, NY, 1980.

[45] Robin Milner, *Communication and Concurrency*, Prentice-Hall International, 1989.

[46] Alfred Müller, *How does the value function of a Markov decision process depend on the transition probabilities?*, Mathematics of Operations Research, 22 (1997), pp. 872–885.

[47] James Orlin, *A faster strongly polynomial minimum cost flow algorithm*, in STOC '88: Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing, New York, NY, USA, 1988, ACM Press, pp. 377–387.

[48] Ronald Ortner, *Pseudometrics for state aggregation in average reward Markov decision processes*, in Algorithmic Learning Theory: 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007, Proceedings Series: Lecture Notes in Computer Science , Vol. 4754 Sublibrary: Lecture Notes in Artificial Intelligence, Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, eds., vol. 4754 of Lecture Notes in Computer Science, Springer-Verlag, 2007, pp. 373–387.

[49] Prakash Panangaden, *Labelled Markov Processes*, Imperial College Press, 2009.

[50] David Park, *Concurrency and automata on infinite sequences*, in Proceedings of the 5th GI-Conference on Theoretical Computer Science, London, UK, 1981, Springer-Verlag, pp. 167–183.

[51] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic, New York, 1967.

[52] Caitlin Phillips, *Knowledge transfer in Markov decision processes*. `http://www.cra.org/Activities/craw/cdmp/awards/2006/Phillips/summary.pdf`, 2006. Final report for Canadian Distributed Mentors Project scholarship.

[53] Martin L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc., New York, NY, USA, 1994.

[54] S.T. Rachev and Ludger Rueschendorf, *Mass transportation problems. Vol. 1: Theory. Vol. 2: Applications.*, Springer Series in Statistics. Probability and its Applications, Springer, New York, 1998.

[55] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, New York, third ed., 1976.

[56] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 1998.

[57] Franck van Breugel, Babita Sharma, and James Worrell, *Approximating a behavioural pseudometric without discount for probabilistic systems*, in Foundations of Software Science and Computational Structures, 10th International Conference, FOSSACS 2007, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2007, Braga, Portugal, March 24-April 1, 2007, Proceedings, Helmut Seidl, ed., vol. 4423 of Lecture Notes in Computer Science, Springer, 2007, pp. 123–137.

[58] Franck van Breugel and James Worrell, *Towards quantitative verification of probabilistic transition systems*, in ICALP '01: Proceedings of the 28th International Colloquium on Automata, Languages and Programming,, London, UK, 2001a, Springer-Verlag, pp. 421–432.

[59] ———, *An algorithm for quantitative verification of probabilistic transition systems*, in CONCUR '01: Proceedings of the 12th International Conference on Concurrency Theory, London, UK, 2001b, Springer-Verlag, pp. 336–350.

[60] Cédric Villani, *Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58)*, American Mathematical Society, 2003.

[61] Jens Vygen, *On dual minimum cost flow algorithms (extended abstract)*, in STOC '00: Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, New York, NY, USA, 2000, ACM, pp. 117–125.

[62] Julian Webster, *Finite approximation of measure and integration.*, Annals of Pure and Applied Logic, 137 (2006), pp. 439–449.

[63] Glynn Winskel, *The Formal Semantics of Programming Languages*, Foundations of Computer Science Series, MITP, Cambridge, Mass., 1993.

[64] Lijun Zhang, Holger Hermanns, Friedrich Eisenbrand, and David N. Jansen, *Flow faster: Efficient decision algorithms for probabilistic simulations*, in Proceedings of the 13th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS'07, Berlin, Heidelberg, 2007, Springer-Verlag, pp. 155–169.