The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Metrics and Continuity in Reinforcement Learning

Charline Le Lan, [*] [1] **Marc G. Bellemare,** [2] **Pablo Samuel Castro**[2]

[1]University of Oxford,
[2]Google Research, Brain Team
charline.lelan@stats.ox.ac.uk, {bellemare,psc}@google.com

Abstract

In most practical applications of reinforcement learning, it is untenable to maintain direct estimates for individual states; in continuous-state systems, it is impossible. Instead, researchers often leverage *state similarity* (whether explicitly or implicitly) to build models that can generalize well from a limited set of samples. The notion of state similarity used, and the neighbourhoods and topologies they induce, is thus of crucial importance, as it will directly affect the performance of the algorithms. Indeed, a number of recent works introduce algorithms assuming the existence of "well-behaved" neighbourhoods, but leave the full specification of such topologies for future work. In this paper we introduce a unified formalism for defining these topologies through the lens of metrics. We establish a hierarchy amongst these metrics and demonstrate their theoretical implications on the Markov Decision Process specifying the reinforcement learning problem. We complement our theoretical results with empirical evaluations showcasing the differences between the metrics considered.

Introduction

A simple principle to generalization in reinforcement learning is to require that similar states be assigned similar predictions. State aggregation implements a coarse version of this principle, by using a notion of similarity to group states together. A finer implementation is to use the similarity in an adaptive fashion, for example by means of a nearest neighbour scheme over representative states. This approach is classically employed in the design of algorithms for continuous state spaces, where the fundamental assumption is the existence of a metric characterizing the real-valued distance between states.

To illustrate this idea, consider the three similarity metrics depicted in Figure 1. The metric d_1 isolates each state, the metric d_3 groups together all states, while the metric d_2 aggregates states based on the similarity in their *long-term dynamics*. In terms of generalization, d_1 would not be expected to generalize well as new states cannot leverage knowledge from previous states; d_3 can cheaply generalize to new states, but at the expense of accuracy; on the other hand, d_2 seems to strike a good balance between the two extremes.

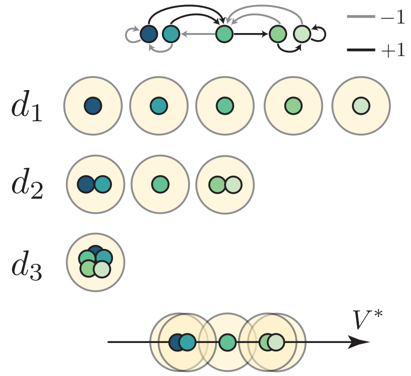

Figure 1: A simple five-state MDP (top) with the neighbourhoods induced by three metrics: an identity metric which isolates each state (d_1); a metric which captures behavioral proximity (d_2); and a metric which is not able to distinguish states (d_3). The yellow circles represent ϵ-balls in the corresponding metric spaces. The bottom row indicates the V^* values for each state.

In this paper we study the effectiveness of *behavioural metrics* at providing a good notion of state similarity. We call behavioural metrics the class of metrics derived from properties of the environment, typically measuring differences in reward and transition functions. Since the introduction of bisimulation metrics (Ferns, Panangaden, and Precup 2004, 2005), a number of behavioural metrics have emerged with additional desirable properties, including lax bisimulation (Taylor, Precup, and Panagaden 2009; Castro and Precup 2010) and π-bisimulation metrics (Castro 2020). Behavioural metrics are of particular interest in the context of understanding generalization, since they directly encode the differences in action-conditional outcomes between states, and hence allow us to make meaningful statements about the relationship between these states.

We focus on the interplay between behavioural metrics and the continuity properties they induce on various functions of interest in reinforcement learning. Returning to our example, V^* is only continuous with respect to d_1 and d_2. The continuity of a set of functions (with respect to a given metric) is assumed in most theoretical results for continuous state

[*]Work performed while a Google Student Researcher.

spaces, such as uniform continuity of the transition function (Kakade, Kearns, and Langford 2003); Lipschitz continuity of all Q-functions of policies (Pazis and Parr 2013), Lipschitz continuity of the rewards and transitions (Zhao and Zhu 2014; Ok, Proutiere, and Tranos 2018) or of the optimal Q-function (Song and Sun 2019; Touati, Taiga, and Bellemare 2020; Sinclair, Banerjee, and Yu 2019). We find that behavioural metrics support these algorithms to varying degrees: the original bisimulation metric, for example, provides fewer guarantees than what is required by some near-optimal exploration algorithms (Pazis and Parr 2013). These results are particularly significant given that behavioural metrics form a relatively privileged group: any metric that enables generalization must in some sense reflect the structure of interactions within the environment and hence, act like a behavioural metric.

Overview

Our aim is to unify representations of state spaces and the notion of continuity via a taxonomy of metrics.

Our first contribution is a general result about the continuity relationships of different functions of the MDP (Theorem 1). While Gelada et al. (2019) (resp. Norets (2010)) proved the uniform Lipschitz continuity of the optimal action-value function (resp. local continuity of the optimal value function) given the uniform Lipschitz continuity (resp. local continuity) of the reward and transition functions and Rachelson and Lagoudakis (2010) showed the uniform Lipschitz continuity of the value function given the uniform Lipschitz continuity of the action-value function in the case of deterministic policies, Theorem 1 is a more comprehensive result about the different components of the MDP (reward and transition functions, value and action value functions), for a spectrum of continuity notions (local and uniform continuity, local and uniform Lipschitz continuity) and applicable with stochastic policies, also providing counterexamples demonstrating that these relationships are only implication results.

Our second contribution is to demonstrate that different metrics lead to different notions of continuity for different classes of functions (Section Continuity: Prior metrics, Section value-based metrics and Table 2). We first study metrics that have been introduced in the literature (presented in Section Prior metrics and abstractions). While Li, Walsh, and Littman (2006) provide a unified treatment of some of these metrics, they do not analyse these abstractions through the lens of continuity. Using our taxonomy, we find that most commonly discussed metrics are actually poorly suited for algorithms that convert representations into values, so we introduce new metrics to overcome this shortcoming (section Value-based metrics). We also analyse the relationships between the topologies induced by all the metrics in our taxonomy (Theorem 2).

Finally, we present an empirical evaluation that supports our taxonomy and shows the importance of the choice of a neighbourhood in reinforcement learning algorithms (section Empirical evaluation).

Background

We consider an agent interacting with an environment, modelled as a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ (Puterman 1994). Here $\mathcal{S}$ is a continuous state space with Borel σ-algebra Σ and $\mathcal{A}$ a discrete set of actions. Denoting $\Delta(X)$ to mean the probability distribution over X, we also have that $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, R_{\text{max}}]$ is the measurable reward function, and $\gamma \in [0, 1)$ is the discount factor. We write $\mathcal{P}_s^a$ to denote the next-state distribution over $\mathcal{S}$ resulting from selecting action a in s and write $\mathcal{R}_s^a$ for the corresponding reward.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from states to distributions over actions, describing a particular way of interacting with the environment. We denote the set of all policies by Π. For any policy $\pi \in \Pi$, the value function $V^\pi(s)$ measures the expected discounted sum of rewards received when starting from state $s \in \mathcal{S}$ and acting according to π:

$$V^\pi(s) := \mathbb{E}\Big[\sum_{t \geq 0} \gamma^t \mathcal{R}_{s_t}^{a_t} \, ; \, s_0 = s, a_t \sim \pi(\cdot \mid s_t)\Big].$$

The maximum attainable value is $V_{\text{max}} := \frac{R_{\text{max}}}{1-\gamma}$. The value function satisfies Bellman's equation:

$$V^\pi(s) = \mathop{\mathbb{E}}_{a \sim \pi(\cdot \mid s)} [\mathcal{R}_s^a + \gamma \mathop{\mathbb{E}}_{s' \sim \mathcal{P}_s^a} V^\pi(s')].$$

The state-action value function or Q-function Q^π describes the expected discounted sum of rewards when action $a \in \mathcal{A}$ is selected from the starting state s, and satisfies the recurrence

$$Q^\pi(s, a) = \mathcal{R}_s^a + \gamma \mathop{\mathbb{E}}_{s' \sim \mathcal{P}_s^a} V^\pi(s').$$

A policy π is said to be optimal if it maximizes the value function at all states:

$$V^\pi(s) = \max_{\pi' \in \Pi} V^{\pi'}(s) \text{ for all } s \in \mathcal{S}.$$

The existence of an optimal policy is guaranteed in both finite and infinite state spaces. We will denote this policy $\pi^* \in \Pi$. The corresponding value function and Q-function are denoted respectively V^* and Q^*.

Metrics, Topologies, and Continuity

We begin by recalling standard definitions regarding metrics and continuity, two concepts central to our work.

Definition 1 (Royden, 1968)**.** *A **metric space** $\langle X, d \rangle$ is a nonempty set X of elements (called points) together with a real-valued function d defined on $X \times X$ such that for all x, y, and z in X: $d(x, y) \geq 0$; $d(x, y) = 0$ if and only if $x = y$; $d(x, y) = d(y, x)$ and $d(x, y) \leq d(x, z) + d(z, y)$. The function d is called a **metric**. A **pseudo-metric** d is a metric with the second condition replaced by the weaker condition $x = y \implies d(x, y) = 0$.*

In what follows, we will often use *metric* to stand for *pseudo-metric* for brevity.

A metric d is useful for our purpose as it quantifies, in a real-valued sense, the relationship between states of the environment. Given a state s, a natural question is: What other states are similar to it? The notion of a *topology* gives a formal answer.

Definition 2 (Sutherland, 2009)**.** *A metric space $\langle X, d\rangle$ induces a **topology** $(X, \mathcal{T}_d)$ defined as the collection of open subsets of X; specifically, the subsets $U \subset X$ that satisfy the property that for each $x \in U$, there exists $\epsilon > 0$ such that the ϵ-**neighbourhood** $B_d(x, \epsilon) = \{y \in X | d(y, x) < \epsilon\} \subset U$. Let $(X, \mathcal{T})$ and $(X, \mathcal{T}')$ be two topologies on the same space X. We say that $\mathcal{T}$ is **coarser** than $\mathcal{T}'$, or equivalently that $\mathcal{T}'$ is **finer** than $\mathcal{T}$, if $\mathcal{T} \subset \mathcal{T}'$.*

Given two similar states under a metric d, we are interested in knowing how functions of these states behave. In the introductory example, we asked specifically: how does the optimal value function behave for similar states? This leads us to the notion of functional continuity. Given $f : X \to Y$ a function between a metric space (X, d_X) and a metric space (Y, d_Y),

- **Local continuity (LC)**: f is locally continuous at $x \in X$ if for any $\epsilon > 0$, there exists a $\delta_{x,\epsilon} > 0$ such that for all $x' \in X$, $d_X(x, x') < \delta_{x,\epsilon} \implies d_Y(f(x), f(x')) < \epsilon$. f is said to be locally continuous on X if it is continuous at every point $x \in X$.
- **Uniform continuity (UC)**: f is uniformly continuous on X when given any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that for all $x, x' \in X, d_X(x, x') < \delta_\epsilon \implies d_Y(f(x), f(x')) < \epsilon$.
- **Local Lipschitz continuity (LLC)**: f is locally Lipschitz continuous at $x \in X$ if there exists $\delta_x > 0, K_x > 0$ such that for all $x', x'' \in B_{d_X}(x, \delta_x)$, $d_Y(f(x'), f(x'')) \leq K_x d_X(x', x'')$.
- **Uniform Lipschitz continuity (ULC)**: f is uniformly Lipschitz continuous if there exist $K > 0$ such that for all $x, x' \in X$ we have $d_Y(f(x), f(x')) \leq K d_X(x, x')$.

The relationship between these different forms of continuity is summarized by the following diagram:

$$\begin{tikzcd} UC \arrow[d] & ULC \arrow[l] \arrow[d] \\ LC & LLC \arrow[l] \end{tikzcd} \tag{1}$$

where an arrow indicates implication; for example, any function that is ULC is also UC.

Here, we are interested in functions of states and state-action pairs. Knowing whether a particular function f possesses some continuity property p under a metric d informs us on how well we can extrapolate the value $f(s)$ to other states; in other words, it informs us on the generalization properties of d.

Prior Metrics and Abstractions

The simplest structure is to associate states to distinct groups, what is often called state aggregation (Bertsekas 2011). This gives rise to an *equivalence relation*, which we interpret as a *discrete pseudo-metric*, that is a metric taking a countable range of values.

Definition 3. *An equivalence relation $E \subseteq X \times X$ induces a **discrete pseudo-metric** e^E where $e^E(x, x') = 0$ if $(x, y) \in E$, and 1 otherwise.*

Throughout the text, we will use e to denote discrete pseudo-metrics. Two extremal examples of metrics are the **identity metric** $e^{\mathbb{I}} : \mathcal{S} \times \mathcal{S} \to \{0, 1\}$, induced by the *identity relation* $\mathbb{I} = \{(s, t) \in \mathcal{S} \times \mathcal{S} | s = t\}$ (e.g. d_1 in Figure 1), and the **trivial metric** $e^{\mathbb{T}} : \mathcal{S} \times \mathcal{S} \to \{0\}$ that collapses all states together (e.g. d_3 in Figure 1).

In-between these extremes, η-abstractions (Li, Walsh, and Littman 2006; Abel, Hershkowitz, and Littman 2017) are functions $\phi : \mathcal{S} \to \hat{\mathcal{S}}$ that aggregates states which are mapped close to each other by a function f. That is, given a threshold $\eta \geq 0$ and $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\phi_{f,\eta}(s) = \phi_{f,\eta}(t) \implies |f(s, a) - f(t, a)| \leq \eta$. We list a few choices for f along with the name of the abstraction we will refer to throughout this text in Table 1.

η-abstractions are defined in terms of a particular function of direct relevance to the agent. However, it is not immediately clear whether these abstractions are descriptive, and, more specifically, the kind of continuity properties they support. An alternative is to relate states based on the outcomes that arise from different choices, starting in these states. These are *bisimulation relations* (Givan, Dean, and Greig 2003).

Definition 4. *An equivalence relation $E \subseteq \mathcal{S} \times \mathcal{S}$ with $\mathcal{S}_E$ the quotient space and $\Sigma(E)$ the Σ measurable sets closed under E, if whenever $(s, t) \in E$ we have:*

- ***Bisimulation relation**[Givan, Dean, and Greig, 2003]. Behavioral indistinguishability under equal actions; namely, for any action $a \in \mathcal{A}$, $\mathcal{R}^a_s = \mathcal{R}^a_t$, and $\mathcal{P}^a_s(X) = \mathcal{P}^a_t(X)$ for all $X \in \Sigma(E)$. We call E a* **bisimulation relation**. *We denote the largest bisimulation relation as $\sim$, and its corresponding discrete metric as $e^{\sim}$.*
- **Lax-bisimulation relation** *[Taylor, Precup, and Panagaden, 2009].*
Behavioral indistinguishability under matching actions; namely, for any action $a \in \mathcal{A}$ from state s there is an action $b \in \mathcal{A}$ from state t such that $\mathcal{R}^a_s = \mathcal{R}^b_t$, and $\mathcal{P}^a_s(X) = \mathcal{P}^b_t(X)$ for all $X \in \Sigma(E)$, and vice-versa, we call E a **lax-bisimulation relation**. *We denote the largest lax-bisimulation relation as $\sim_{lax}$, and its corresponding discrete metric as $e^{\sim_{lax}}$.*
- **π-bisimulation relation** *[Castro, 2020]. Behavioral indistinguishability under a fixed policy; namely, given a policy $\pi \in \Pi$, $\sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}^a_s = \sum_{a \in \mathcal{A}} \pi(a|t) \mathcal{R}^a_t$, and $\sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^a_s(X) = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^b_t(X)$ for all $X \in \Sigma(E)$. We call E a* **π-bisimulation relation**. *We denote the largest bisimulation relation as $\sim_\pi$, and its corresponding discrete metric as $e^{\sim_\pi}$.*

A *bisimulation metric* is the continous generalization of a bisimulation relation. Formally, d is a bisimulation metric if its kernel is equivalent to the bisimulation relation. The canonical bisimulation metric (Ferns, Panangaden, and Precup 2005) is constructed from the Wasserstein distance between probability distributions.

Definition 5. *Let (Y, d_Y) be a metric space with Borel σ-algebra Σ. The Wasserstein distance (Villani 2008) between two probability measures P and Q on Y, under a given metric d_Y is given by $W_{d_Y}(P, Q) =$*

f	$\phi_{f,\eta}$
Q^*	approximate Q function abstraction ($\eta \geq 0$) / Q^*-irrelavance ($\eta = 0$)
$\mathcal{R}$ and $\mathcal{P}$	approximate model abstraction ($\eta \geq 0$) / Model-irrelevance ($\eta = 0$)
Q^π	Q^π-irrelevance abstraction ($\eta = 0$)
$\max\limits_{\mathcal{A}} Q^*$	a^*-irrelevance abstraction ($\eta = 0$)

Table 1: Different types of state abstractions.

$\inf_{\lambda \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\lambda}[d_Y(x,y)]$, where $\Gamma(P,Q)$ is the set of couplings between P and Q.

Lemma 1 (Ferns, Panangaden, and Precup, 2005)**.** *Let $\mathcal{M}$ be the space of state pseudo-metrics and define the functional $F : \mathcal{M} \to \mathcal{M}$ as $F(d)(x,y) = \max_{a\in\mathcal{A}} \left(|\mathcal{R}_x^a - \mathcal{R}_y^a| + \gamma W_d(\mathcal{P}_x^a, \mathcal{P}_y^a)\right)$. Then F has a least fixed point $d^\sim$ and $d^\sim$ is a bisimulation metric.*

In words, bisimulation metrics arise as the fixed points of an operator on the space of pseudo-metrics. Lax bisimulation metrics $d^{\sim_{lax}}$ and a π-bisimulation metrics $d^{\sim_\pi}$ can be defined in an analogous fashion; for succinctness, their formal definitions are included in the appendix.

Continuity Relationships

Our first result characterizes the continuity relationships between key functions of the MDP. The theorem considers different forms of continuity and relates how the continuity of one function implies another. While the particular case of uniform Lipschitz continuity of Q^* (resp. local continuity of V^*) from $\mathcal{P} + \mathcal{R}$ has been remarked on before by Gelada et al. (2019) (resp. Norets (2010)) as well as the case of uniform Lipschitz continuity of V^π given the uniform Lipschitz continuity of Q^π for stochastic policies π (Rachelson and Lagoudakis 2010), to the best of our knowledge this is the first comprehensive treatment of the topic, in particular providing counterexamples.

Theorem 1. *If we decompose the Cartesian product $\mathcal{S} \times \mathcal{A}$ as: $d_{S\times A}(s,a,s',a') = d_S(s,s') + d_A(a,a')$ with $d_{\mathcal{A}}$ the identity metric, the LC, UC and LLC relationships between $\mathcal{P}$, $\mathcal{R}$, V^π, V^*, Q^π and Q^* functions are given by diagram 3. A directed arrow $f \to g$ indicates that function g is continuous whenever f is continuous. Labels on arrows indicate conditions that are necessary for that implication to hold. $\mathcal{P}+\mathcal{R}$ is meant to stand for both $\mathcal{P}$ and $\mathcal{R}$ continuity; π-cont indicates continuity of $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. An absence of a directed arrow indicates that there exists a counter-example proving that the implication does not exist. In the ULC case, the previous relationships also hold with the following additional assumptions: $\gamma L_{\mathcal{P}} < 1$ for $\mathcal{P} + \mathcal{R} \to Q^*$ and $\gamma L_{\mathcal{P}}(1 + L_\pi) < 1$ for $\mathcal{P} + \mathcal{R} \xrightarrow{\pi\text{-cont}} Q^\pi$ where $L_{\mathcal{P}}$ and L_π are the Lipschitz constants of $\mathcal{P}$ and π, respectively.*

$$\begin{tikzcd}
 & Q^\pi \arrow[r, "\pi\text{-cont}"] & V^\pi \\
\mathcal{P}+\mathcal{R} \arrow[ur, "\pi\text{-cont}"] \arrow[dr] & & \\
 & Q^* \arrow[r] & V^*
\end{tikzcd} \tag{2}$$

Proof. All proofs and counterexamples are provided in the appendix. □

The arrows are transitive and apply for all forms of continuity illustrated in Diagram (1); for example, if we have ULC for Q^*, this implies we have LC for V^*. This diagram is useful when evaluating metrics as they clarify the strongest (or weakest) form of continuity one can demonstrate. When considering deterministic policies, we can notice that the π-continuity mentioned in Theorem 1 is very restrictive, as the following lemma shows.

Lemma 2. *If a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is continuous, $\mathcal{S}$ is connected[1] and $\mathcal{A}$ is discrete, then π is globally constant.*

Taxonomy of Metrics

We now study how different metrics support the continuity of functions relevant to reinforcement learning and the relationship between their induced topologies. While the taxonomy we present here is of independent interest, it also provides a clear theoretical foundation on which to build results regarding metric-respecting embeddings (Gelada et al. 2019; Zhang et al. 2020).

Continuity: Prior Metrics

We begin the exposition by considering the continuity induced by discrete metrics. These enable us to analyze the properties of some representations found in the literature. The extremes of our metric hierarchy are the identity metric $e^{\mathbb{I}}$ and trivial metric $e^{\mathbb{T}}$, which respectively support all and one continuous functions, and were represented by d_1 and d_3 in the introductory example.

Lemma 3 (Identity metric)**.** *$e^{\mathbb{I}}$ induces the finest topology on $\mathcal{S}$, made of all possible subsets of $\mathcal{S}$. Let (Y, d_Y) be any metric space. Any function h (resp. Any bounded h) $: (\mathcal{S}, e^{\mathbb{I}}) \to (Y, d_Y)$ is LC and UC (resp. ULC).*

[1] A connected space is topological space that cannot be represented as the union of two or more disjoint non-empty open subsets.

Lemma 4 (Trivial metric). *$e^{\mathbb{T}}$ induces the coarsest topology on $\mathcal{S}$, consisting solely of $\{\emptyset, \mathcal{S}\}$. Let (Y, d_Y) be any metric space. Any function $h : (\mathcal{S}, e^{\mathbb{I}}) \to (Y, d_Y)$ is LC, UC and ULC iff h is constant.*

We can also construct a discrete metric from any state aggregation $\phi : \mathcal{S} \to \hat{\mathcal{S}}$ as $e^{\phi}(s,t) = e^{\mathbb{I}}(\phi(s), \phi(t)) = 0$ if $\phi(s) = \phi(t)$, and 1 otherwise. However, as stated below, η-abstractions do not guarantee continuity except in the trivial case where $\eta = 0$.

Lemma 5. *If $\eta = 0$, then any function f (resp. bounded function f): $(\mathcal{S}, d_{\mathcal{S}}) \to (Y, d_Y)$ is LC and UC (resp. ULC) with respect to the pseudometric $e^{\phi_{f,\eta}}$. However, given a function f and $\eta > 0$, there exists an η-abstraction $\phi_{f,\eta}$ such that f is not continuous with respect to $e^{\phi_{f,\eta}}$.*

Unlike the discrete metrics defined by η-abstractions, both bisimulation metrics and the metric induced by the bisimulation relation support continuity of the optimal value function.

Lemma 6. *Q^* (resp. Q^π) is ULC with Lipschitz constant 1 with respect to $d^{\sim}$ (resp. $d^{\sim_\pi}$).*

Corollary 1. *Q^* (resp. Q^π) is ULC with Lipschitz constant V_{max} with respect to $e^{\sim}$ (resp. $e^{\sim_\pi}$).*

We note that Ferns, Panangaden, and Precup (2004) proved a weaker statement involving V^* (resp. Castro, Panangaden, and Precup (2009), V^π). To summarize, metrics that are too coarse may fail to provide the requisite continuity of reinforcement learning functions. Bisimulation metrics are particularly desirable as they achieve both a certain degree of coarseness, while preserving continuity. In practice, however, Ferns, Panangaden, and Precup's bisimulation metric is difficult to compute and estimate, and tends to be conservative – as long as two states can be distinguished by action sequences, bisimulation will keep them apart.

Value-Based Metrics

As an alternative to bisimulation metrics, we consider simple metrics constructed from value functions and study their continuity offerings. These metrics are simple in that they are defined in terms of differences between values, or functions of values, at the states being compared. The last metric, $d_{\Delta_\forall}$, is particularly appealing as it can be approximated, as we describe below. Under this metric, all Q-functions are Lipschitz continuous, supporting some of the more demanding continuous-state exploration algorithms (Pazis and Parr 2013).

Lemma 7. *For a given MDP, let Q^π be the Q-function of policy π, and Q^* the optimal Q-function. The following are continuous pseudo-metrics:*

1. $d_{\Delta^*}(s,s') = \max_{a \in \mathcal{A}} |Q^*(s,a) - Q^*(s',a)|$
2. $d_{\Delta_\pi}(s,s') = \max_{a \in \mathcal{A}} |Q^\pi(s,a) - Q^\pi(s',a)|$
3. $d_{\Delta_\forall}(s,s') = \max_{\pi \in \Pi, a \in \mathcal{A}} |Q^\pi(s,a) - Q^\pi(s',a)|$

Q^* (resp. Q^π) is ULC with Lipschitz constant 1 wrt to d_{Δ^*} (resp. d_{Δ^π}). Q^π is ULC with Lipschitz constant 1 wrt to $d_{\Delta_\forall}$ for any $\pi \in \Pi$.

Remark. *When $\mathcal{S}$ is finite, the number of policies to consider to compute $d_{\Delta_\forall}$ is finite: $d_{\Delta_\forall}(s,s') = \max_{\pi \in \Pi, a \in \mathcal{A}} |Q^\pi(s,a) - Q^\pi(s',a)| = \max_{\pi \in \Pi_{AVF}, a \in \mathcal{A}} |Q^\pi(s,a) - Q^\pi(s',a)|$, where Π_{AVF} is the finite set of extremal policies corresponding to Adversarial Value Functions (AVFs) (Bellemare et al. 2019).*

$d_{\Delta_\forall}$ provides strong continuity of the value-function for all policies contrary to any other metric that has been used in the literature. Since computing $d_{\Delta_\forall}$ is computationally expensive, we will approximate it by the pseudometric $d_{\widetilde{\text{AVF}}(n)} = \max_{\pi \in \Pi_{\widetilde{\text{AVF}}(n)}, a \in \mathcal{A}} |Q^\pi(s,a) - Q^\pi(s',a)|$, where $\Pi_{\widetilde{\text{AVF}}(n)}$ are n samples from the set of extremal policies Π_{AVF}.

Categorizing Metrics, Continuity and Complexity

We now formally present in Theorem 2 the topological relationships between the different metrics. This hierarchy is important for generalization purposes as it provides a comparison between the shapes of different neighbourhoods which serve as a basis for RL algorithms on continuous state spaces.

Theorem 2. *The relationships between the topologies induced by the metrics in Table 2 are given by the following diagram. We denote by $d_1 \to d_2$ when $\mathcal{T}_{d_1} \subset \mathcal{T}_{d_2}$, that is, when $\mathcal{T}_{d_1}$ is coarser than $\mathcal{T}_{d_2}$. Here d denotes any arbitrary metric.*

$$\begin{tikzcd}
e^{\sim_{lax}} \arrow[d] & d^{\sim_{lax}} \arrow[l] \arrow[d] & d_{\widetilde{AVF}(n)} \arrow[dr] & d \arrow[r] & e^{\mathbb{I}} \\
e^{\sim} & d^{\sim} \arrow[l] & d_{\Delta^*} \arrow[l] \arrow[r] & d_{\Delta_\forall} & d_{\Delta_\pi} \arrow[l] \arrow[d] \\
d & e^{\mathbb{T}} \arrow[l] & & e^{\sim_\pi} & d^{\sim_\pi} \arrow[l]
\end{tikzcd}$$

Proof. All proofs can be found in the appendix. The relation $d^{\sim_{lax}} \to d^{\sim}$ was shown by Taylor, Precup, and Panagaden (2009) but not expressed in topological terms. □

We summarize in Table 2 our continuity results mentioned throughout this section and supplement them with the continuity of the lax-bisimulation metric proven in Taylor, Precup, and Panagaden (2009). To avoid over-cluttering the table, we only specify the strongest form of functional continuity according to Theorem 1. As an additional key differentiator, we also note the complexity of computing these metrics from a full model of the environment, which gives some indication about the difficulty of performing state abstraction. Proofs are provided in the appendix.

From a computational point of view, all continuous metrics can be approximated using deep learning techniques which makes them even more attractive to build representations. Atari 2600 experiments by Castro (2020) show that π-bisimulation metrics do perform well in larger domains. This is also supported by (Zhang et al. 2020) who use an encoder architecture to learn a representation that respects the bisimulation metric.

Metric	LC	UC	ULC	LLC	Complexity
Discrete metric $e^{\mathbb{I}}$	$Y^{\mathcal{S}}$	$Y^{\mathcal{S}}$	$\mathcal{B}(Y^{\mathcal{S}})$	$\mathcal{B}_L(Y^{\mathcal{S}})$	$O(\lvert\mathcal{S}\rvert)$
Trivial metric $e^{\mathbb{T}}$	$\{y\}^{\mathcal{S}}$	$\{y\}^{\mathcal{S}}$	$\{y\}^{\mathcal{S}}$	$\{y\}^{\mathcal{S}}$	$O(1)$
Model-irrelevance	$\mathcal{P}, \mathcal{R}$	$\mathcal{P}, \mathcal{R}$	$\mathcal{P}, \mathcal{R}$	$\mathcal{P}, \mathcal{R}$	
Q^π-irrelevance	Q^π	Q^π	Q^π	Q^π	
Q^*-irrelevance	Q^*	Q^*	Q^*	Q^*	
a^*-irrelevance	Q^*	Q^*	Q^*	Q^*	
Approx. abstraction	-	-	-	-	
$e^{\sim}$	Q^*	Q^*	Q^*	Q^*	$O(\lvert\mathcal{A}\rvert\lvert\mathcal{S}\rvert^3)$
$d^{\sim}$	Q^*	Q^*	Q^*	Q^*	$O\big(\lvert\mathcal{A}\rvert\lvert\mathcal{S}\rvert^5 \log\lvert\mathcal{S}\rvert \frac{\ln\delta}{\ln\gamma}\big)$
$e^{\sim_\pi}$	Q^π	Q^π	Q^π	Q^π	$O(\lvert\mathcal{S}\rvert^3)$
$d^{\sim_\pi}$	Q^π	Q^π	Q^π	Q^π	$O\big(\lvert\mathcal{S}\rvert^5 \log\lvert\mathcal{S}\rvert \frac{\ln\delta}{\ln\gamma}\big)$
$e^{\sim_{lax}}$	V^*	V^*	V^*	V^*	$O(\lvert\mathcal{A}\rvert^2\lvert\mathcal{S}\rvert^3)$
$d^{\sim_{lax}}$	V^*	V^*	V^*	V^*	$O\big(\lvert\mathcal{A}\rvert^2\lvert\mathcal{S}\rvert^5 \log\lvert\mathcal{S}\rvert \frac{\ln\delta}{\ln\gamma}\big)$
d_{Δ^*}	Q^*	Q^*	Q^*	Q^*	$O\big(\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert \frac{\log(\mathcal{R}_{\max}^{-1}\delta(1-\gamma))}{\log(\gamma)}\big)$
d_{Δ_π}	Q^π	Q^π	Q^π	Q^π	$O\big(\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert \frac{\log(\mathcal{R}_{\max}^{-1}\delta(1-\gamma))}{\log(\gamma)}\big)$
$d_{\Delta_\forall}$	$Q^\pi, \forall\pi \in \Pi$	$Q^\pi, \forall\pi \in \Pi$	$Q^\pi, \forall\pi \in \Pi$	$Q^\pi, \forall\pi \in \Pi$	NP-hard? (Bellemare et al. 2019)

Table 2: Categorization of state metrics, their continuity implications, and their complexity (when known). The notation $\{y\}^{\mathcal{S}}$ denotes any function $h : \mathcal{S} \to Y$ that is constant, $Y^{\mathcal{S}}$ refers to all functions $h : \mathcal{S} \to Y$. $\mathcal{B}(Y^{\mathcal{S}})$ (resp. $\mathcal{B}_L(Y^{\mathcal{S}})$) is a bounded (resp. locally bounded) function $h : \mathcal{S} \to Y$. "-" denotes an absence of LC, UC, ULC and LLC. In the complexity column, δ is the desired accuracy.

Empirical Evaluation

We now conduct an empirical evaluation to quantify the magnitude of the effects studied in the previous sections. Specifically, we are interested in how approximations derived from different metrics impact the performance of basic reinforcement learning procedures. We consider two kinds of approximations: state aggregation and nearest neighbour, which we combine with six representative metrics: $e^{\sim}$, $e^{\sim_{lax}}$, $d^{\sim}$, $d^{\sim_{lax}}$, d_{Δ^*}, and $d_{\widetilde{\text{AVF}(50)}}$.

We conduct our experiments on Garnet MDPs, which are a class of randomly generated MDPs (Archibald, McKinnon, and Thomas 1995; Piot, Geist, and Pietquin 2014). Specifically, a Garnet MDP $Garnet(n_{\mathcal{S}}, n_{\mathcal{A}})$ is parameterized by two values: the number of states $n_{\mathcal{S}}$ and the number of actions $n_{\mathcal{A}}$, and is generated as follows: **1.** The branching factor $b_{s,a}$ of each transition $\mathcal{P}_s^a$ is sampled uniformly from $[1 : n_{\mathcal{S}}]$. **2.** $b_{s,a}$ states are picked uniformly randomly from $\mathcal{S}$ and assigned a random value in $[0, 1]$; these values are then normalized to produce a proper distribution $\mathcal{P}_s^a$. **3.** Each $\mathcal{R}_s^a$ is sampled uniformly in $[0, 1]$. The use of Garnet MDPs grants us a less-biased comparison of the different metrics than if we were to pick a few specific MDPs. Nonetheless, we do provide extra experiments on a set of GridWorld tasks in the appendix. The code used to produce all these experiments is open-sourced [2].

Generalizing the Value Function V^*

We begin by studying the approximation error that arises when extrapolating the optimal value function V^* from a subset of states. Specifically, given a subsampling fraction $f \in [0, 1]$, we sample $\lceil|\mathcal{S}| \times f\rceil$ states and call this set κ. For each unknown state $s \in \mathcal{S} \setminus \kappa$, we find its nearest known neighbour according to metric d: $NN(s) = \arg\min_{t\in\kappa} d(s, t)$. We then define the optimal value function as $\hat{V}^*(s) = V^*(NN(s))$, and report the approximation error in Figure 2 (left). This experiment gives us insights into how amenable the different metrics are for transferring value estimates across states; effectively, their generalization capabilities.

According to Theorem 2, the two discrete metrics $e^{\sim}$ and $e^{\sim_{lax}}$ induce finer topologies than their four continuous counterparts. Most of the states being isolated from each other in these two representations, $e^{\sim}$ and $e^{\sim_{lax}}$ perform poorly. The three continuous metrics $d^{\sim}$, $d^{\sim_{lax}}$ and d_{Δ^*} all guarantee Lipschitz continuity of V^* while $d_{\widetilde{\text{AVF}(50)}}$ is approximately V^* Lipschitz continuous. However, $d^{\sim_{lax}}$ (resp. d_{Δ^*}) produce coarser (resp. approximately coarser) topologies than $d^{\sim}$ (resp. $d_{\widetilde{\text{AVF}(50)}}$) (see Theorem 2). This is reflected in their better generalization error compared to the latter two metrics. Additionally, the lax bisimulation metric $d^{\sim_{lax}}$ outperforms d_{Δ^*} substantially, which can be explained by noting that $d^{\sim_{lax}}$ measures distances between two states under independent action choices, contrary to all other metrics.

Generalizing the Q-function Q^*

We now illustrate the continuity (or absence thereof) of Q^* with respect to the different metrics. In Figure 2 (center), we perform a similar experiment as the previous one, still using a 1-nearest neighbour scheme but now extrapolating Q^*.

As expected, we find that metrics that do not support Q^*

[2] Code available at https://github.com/google-research/google-research/tree/master/rl_metrics_aaai2021

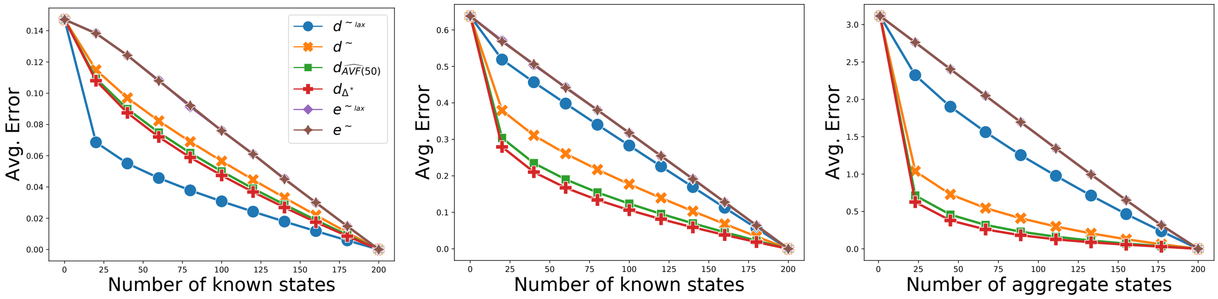

Figure 2: Errors when approximating the optimal value function (left) and optimal Q-function (center) via nearest-neighbours and errors when performing value iteration on aggregated states (right). Curves for $e^{\sim}$ and $e^{\sim_{lax}}$ are covering each other on all of the plots. Averaged over 100 Garnet MDPs with 200 states and 5 actions, with 50 independent runs for each (to account for subsampling differences). Confidence intervals were very tiny due to the large number of runs so were not included.

continuity, including $d^{\sim_{lax}}$, cannot generalize from a subset of states, and their average error decreases linearly. In contrast, the three other metrics are able to generalize. Naturally, d_{Δ^*}, which aggregates states based on Q^*, performs particularly well. However, we note that $d_{\Delta_\forall}$ also outperforms the bisimulation metric $d^{\sim}$, highlighting the latter's conservativeness, which tends to separate states more. By our earlier argument regarding $d^{\sim_{lax}}$, this suggests there may be a class of functions, not represented in Table 2, which is continuous under $d^{\sim}$ but not $d_{\widetilde{\text{AVF}}(50)}$.

Approximate Value Iteration

As a final experiment, we perform approximate value iteration using a state aggregation ϕ derived from one of the metrics. For each metric, we perform 10 different aggregations using a k-median algorithm, ranging from one aggregate state to 200 aggregate states. For a given aggregate state c, let $Q(c, a)$ stand for its associated Q-value. The approximation value iteration update is

$$\hat{Q}_k(c, a) \leftarrow \frac{1}{|c|} \sum_{s|\phi(s)=c} \left[\mathcal{R}_s^a + \gamma \mathbb{E}_{s' \sim \mathcal{P}_s^a} \max_{a \in \mathcal{A}} \hat{Q}_k(\phi(s')) \right]$$

We can then measure the error induced by our aggregation via $\max_{a \in \mathcal{A}} \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} |Q^*(s, a) - \hat{Q}_k(\phi(s), a)|$, which we display in the rightmost panel of Figure 2.

As in our second experiment, the metrics that do not support Q^*-continuity well fail to give good abstractions for approximate value iteration. As for $e^{\sim}$, the topology induced by this metric is too fine (Theorem 2) leading to poor generalisation results. The performance of d_{Δ^*} is consistent with Theorem 2, which states that it induces the coarsest topology. However, although it is known that Q^*-continuity is sufficient for approximate value iteration (Li, Walsh, and Littman 2006), it is somewhat surprising that it outperforms $d_{\widetilde{\text{AVF}}(50)}$, since $d_{\widetilde{\text{AVF}}(50)}$ is an approximant of $d_{\Delta_\forall}$ that is designed to provide continuity with respect to all policies, so it may be expected to yield better approximations at intermediate iterations. Despite this, $d_{\Delta_\forall}$ still serves as an interesting and tractable surrogate metric to d_{Δ^*}.

Discussion

Behavioral metrics are important both to evaluate the goodness of a given state representation and to learn such a representation. We saw that approximate abstractions and equivalence relations are insufficient for continuous-state RL problems, because they do not support the continuity of common RL functions or induce very fine representations on the state space leading to poor generalization.

Continuous behavioural metrics go one step further by considering the structure of the MDP in their construction and inducing coarser topologies than their discrete counterparts; however, within that class we still find that not all metrics are equally useful. The original bisimulation metric of Ferns, Panangaden, and Precup (2004), for example, is too conservative and has a rather fine topology. This is confirmed by our experiments in Figure 2, where it performs poorly overall. The lax bisimulation metric guarantees the continuity of V^* which makes it suitable for transferring optimal values between states but fails to preserve continuity of Q^*. Together with our analysis, the d_{Δ^*} and $d_{\Delta_\forall}$ metrics seem interesting candidates when generalising within a neighbourhood.

$d_{\Delta_\forall}$ is useful when we do not know the value improvement path the algorithm will be following (Dabney et al. 2020). Despite being approximated from a finite number of policies, the performance of $d_{\widetilde{\text{AVF}}(n)}$, reflects the fact that it respects, in some sense, the entire space of policies that are spanned by policy iteration and makes it useful in practice. One advantage of this metric is that it is built from value functions, which are defined on a per-state basis; this makes it amenable to online approximations. In contrast, bisimulation metrics are only defined for pairs of states, which makes it difficult to approximate in an online fashion, specifically due to the difficulty of estimating the Wasserstein metric on every update.

Finally, continuing our analysis on partially observable systems is an interesting area for future work. Although Castro, Panangaden, and Precup (2009) proposed various equivalence relations for partially observable systems, there has been little work in defining proper metrics for these systems.

Acknowledgements

The authors would like to thank Sheheryar Zaidi, Adam Foster and Abe Ng for insightful discussions about topology and functional analysis, Carles Gelada, John D. Martin, Dibya Ghosh, Ahmed Touati, Rishabh Agarwal, Marlos Machado and the whole Google Brain team in Montreal for helpful discussions, and Robert Dadashi for a conversation about polytopes. We also thank Lihong Li and the anonymous reviewers for useful feedback on this paper.

We would also like to thank the Python community (Van Rossum and Drake Jr 1995; Oliphant 2007) and in particular NumPy (Oliphant 2006; Walt, Colbert, and Varoquaux 2011; Harris et al. 2020), Tensorflow (Abadi et al. 2016), SciPy (Jones et al. 2001), Matplotlib (Hunter 2007) and Gin-Config (https://github.com/google/gin-configGin-Config).

Broader Impact

This work lies in the realm of "foundational RL" in that it contributes to the fundamental understanding and development of reinforcement learning algorithms and theory. As such, despite us agreeing in the importance of this discussion, our work is quite far removed from ethical issues and potential societal consequences.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 265–283.

Abel, D.; Hershkowitz, D. E.; and Littman, M. L. 2017. Near optimal behavior via approximate state abstraction. *arXiv preprint arXiv:1701.04113* .

Archibald, T. W.; McKinnon, K. I. M.; and Thomas, L. C. 1995. On the Generation of Markov Decision Processes. *The Journal of the Operational Research Society* 46(3): 354–361.

Bellemare, M.; Dabney, W.; Dadashi, R.; Ali Taiga, A.; Castro, P. S.; Le Roux, N.; Schuurmans, D.; Lattimore, T.; and Lyle, C. 2019. A Geometric Perspective on Optimal Representations for Reinforcement Learning. In *Advances in Neural Information Processing Systems 32*, 4360–4371.

Bertsekas, D. P. 2011. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications* 9(3): 310–335.

Castro, P. S. 2020. Scalable methods for computing state similarity in deterministic Markov Decision Processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Castro, P. S.; Panangaden, P.; and Precup, D. 2009. Notions of state equivalence under partial observability. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*.

Castro, P. S.; and Precup, D. 2010. Using bisimulation for policy transfer in MDPs. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Dabney, W.; Barreto, A.; Rowland, M.; Dadashi, R.; Quan, J.; Bellemare, M. G.; and Silver, D. 2020. The Value-Improvement Path: Towards Better Representations for Reinforcement Learning. *arXiv preprint arXiv:2006.02243* .

Dadashi, R.; Taïga, A. A.; Roux, N. L.; Schuurmans, D.; and Bellemare, M. G. 2019. The value function polytope in reinforcement learning. *arXiv preprint arXiv:1901.11524* .

Ferns, N.; Panangaden, P.; and Precup, D. 2004. Metrics for finite Markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 162–169. AUAI Press.

Ferns, N.; Panangaden, P.; and Precup, D. 2005. Metrics for Markov Decision Processes with Infinite State Spaces. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*.

Gelada, C.; Kumar, S.; Buckman, J.; Nachum, O.; and Bellemare, M. G. 2019. DeepMDP: Learning Continuous Latent Space Models for Representation Learning. In *Proceedings of the International Conference on Machine Learning*.

Givan, R.; Dean, T.; and Greig, M. 2003. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147(1-2): 163–223.

Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; et al. 2020. Array programming with NumPy. *Nature* 585(7825): 357–362.

Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering* 9(3): 90–95.

Jones, E.; Oliphant, T.; Peterson, P.; et al. 2001. SciPy: Open source scientific tools for Python . http://www.scipy.org/

Kakade, S.; Kearns, M. J.; and Langford, J. 2003. Exploration in metric state spaces. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 306–312.

Li, L.; Walsh, T. J.; and Littman, M. L. 2006. Towards a Unified Theory of State Abstraction for MDPs. In *ISAIM*.

Machado, M.; Bellemare, M.; and Bowling, M. 2017. A Laplacian framework for option discovery in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.

Norets, A. 2010. Continuity and differentiability of expected value functions in dynamic discrete choice models. *Quantitative economics* 1(2): 305–322.

Ok, J.; Proutiere, A.; and Tranos, D. 2018. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, 8874–8882.

Oliphant, T. E. 2006. *A guide to NumPy*, volume 1. Trelgol Publishing USA.

Oliphant, T. E. 2007. Python for scientific computing. *Computing in Science & Engineering* 9(3): 10–20.

Pazis, J.; and Parr, R. 2013. PAC optimal exploration in continuous space Markov decision processes. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Piot, B.; Geist, M.; and Pietquin, O. 2014. Difference of Convex Functions Programming for Reinforcement Learning. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*, 2519–2527. Curran Associates, Inc.

Puterman, M. L. 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming . John Wiley & Sons, Inc.

Rachelson, E. and Lagoudakis, M. G. On the locality of action domination in sequential decision making. In *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2010, Fort Lauderdale, Florida, USA, January 6-8, 2010*, 2010.

Royden, H. 1968. *Real Analysis*. Upper Saddle River, New Jersey 07458: Prentice Hall, 3 edition. ISBN 0024041513.

Sinclair, S. R.; Banerjee, S.; and Yu, C. L. 2019. Adaptive Discretization for Episodic Reinforcement Learning in Metric Spaces. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3(3): 1–44.

Solway, A.; Diuk, C.; Córdova, N.; Yee, D.; Barto, A. G.; Niv, Y.; and Botvinick, M. M. 2014. Optimal Behavioral Hierarchy. *PLoS Computational Biology* 10(8): e1003779.

Song, Z.; and Sun, W. 2019. Efficient model-free reinforcement learning in metric spaces. *arXiv preprint arXiv:1905.00475* .

Sutherland, W. A. 2009. *Introduction to metric and topological spaces*. Oxford University Press.

Sutton, R.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence* 112: 181–211.

Taylor, J.; Precup, D.; and Panagaden, P. 2009. Bounding performance loss in approximate MDP homomorphisms. In *Advances in Neural Information Processing Systems*, 1649–1656.

Touati, A.; Taiga, A. A.; and Bellemare, M. G. 2020. Zooming for Efficient Model-Free Reinforcement Learning in Metric Spaces. *arXiv preprint arXiv:2003.04069* .

Van Rossum, G.; and Drake Jr, F. L. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Walt, S. v. d.; Colbert, S. C.; and Varoquaux, G. 2011. The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering* 13(2): 22–30.

Zhang, A.; McAllister, R.; Calandra, R.; Gal, Y.; and Levine, S. 2020. Learning Invariant Representations for Reinforcement Learning without Reconstruction. *arXiv preprint arXiv:2006.10742* .

Zhao, D.; and Zhu, Y. 2014. MEC—A near-optimal online reinforcement learning algorithm for continuous deterministic systems. *IEEE transactions on neural networks and learning systems* 26(2): 346–356.