State similarity metrics in reinforcement learning

Tyler Kastner

School of Computer Science

McGill University, Montreal

June 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Computer Science

Abstract

Bisimulation relations and metrics are a notion of similarity between states which have originated in the study of transition systems and which have since been extended to probabilistic systems, then to Markov processes and finally Markov decision processes. In this thesis, we first introduce a relaxation of bisimulation metrics, named the MICo distance, which, unlike bisimulation, can be estimated through samples and we analyze its theoretical properties. We study variations of it and prove convergence results related to learning the distance through samples. We then present empirical results obtained from large-scale experiments. We introduce a related quantity, kernel similarity metrics, which are a modification of the original bisimulation metric obtained by iterating over the space of positive definite kernels, rather than over the space of metrics. We analyze various properties of these distances using tools from reproducing kernel Hilbert space theory, and prove that a distance arising from this theory is equivalent to the MICo distance introduced earlier. Finally, we extend and formalize a framework of auxiliary Markov decision processes, which allows us to view the original bisimulation metric and its various modifications in a single framework.

Abrégé

Les relations et les métriques de bisimulation sont une notion de similarité entre états qui trouve son origine dans l'étude des processus de Markov étiquetés, qui ont été étendues aux processus de décision de Markov. Dans cette thèse, nous introduisons d'abord une relaxation des métriques de bisimulation, nommée la distance MICo, qui contrairement à la bisimulation, peut être estimée à travers des échantillons et nous analysons ses propriétés théoriques. Nous en étudions les variations et prouvons des résultats de convergence liés à l'apprentissage de la distance à travers des échantillons. Nous présentons ensuite les résultats empiriques robustes qu'il a obtenu dans des expériences à grande échelle. Nous introduisons une quantité connexe, les métriques de similarité du noyau, qui sont une modification de la métrique de bisimulation originale résultant de l'itération sur l'espace des noyaux définis positifs, plutôt que sur l'espace des métriques. Nous analysons diverses propriétés de ces distances à l'aide d'outils de reproduction de la théorie spatiale du noyau de Hilbert et prouvons une équivalence à la distance MICo introduite précédemment. Enfin, nous étendons et formalisons un cadre théorique des processus de décision de Markov auxiliaires qui unifie les modifications précédentes avec la métrique de bisimulation originale.

Acknowledgements

It is safe to say that without all of the help I've received along the way, none of this would be possible, and I wholeheartedly thank all who have supported me along the way.

Firstly, I must thank Prakash - I could not have asked for a better supervisor. He has guided me through every step of my journey, since my first summer research as an undergraduate, and because of him I have grown tremendously as a researcher. He is an inspiration, and I am honoured to have the opportunity to work under him these past 2 years.

I would also like to thank everyone at the School of Computer Science - the professors who taught me so much, as well as the administrative staff, who helped me through various setbacks and ensured I didn't stray too far from the path.

Most of all, I must thank my family - my parents Coleen and Irwin, my sister Jenna, and my partner Sierra. They have been by my side through all of the ups and downs, and have provided neverending support. I could not have done this without them.

Contribution of authors

Most of chapter 3 appears in the NeurIPS 2021 paper:

- Castro, P. S., Kastner, T., Panangaden, P., and Rowland, M. Mico: Learning improved representations via sampling-based state similarity for Markov decision processes. In Advances in Neural Information Processing Systems (NeurIPS 2021), 2021.

Chapters 4, 5, and 6 contain novel material not yet published. All chapters are joint work between the author, Prakash Panangaden, Mark Rowland, and Pablo Samuel Castro. The main theoretical development was carried out by the author of this thesis with help and guidance from Pablo Castro, Mark Rowland and Prakash Panangaden. The experimental work reported is primarily due to Pablo Castro.

Table of Contents

List of Figures

Chapter 1

Introduction

Reinforcement learning is a field of machine learning concerned with designing agents which learn to make decisions sequentially. The quality of the decisions made is dictated by an extrinsic reward signal, provided to the agent after each decision is made.

The framework most commonly used to describe this setting is the discounted Markov decision process (MDP). It consists of a tuple of a state space, action space, transition function, reward function, and discount factor. The state space describes the possible states that the environment is in, from which an agent must make decisions. The action space represents the actions that an agent can make, these choices are the decisions the agent learns. The transition and reward functions dictate how the environment dynamics evolve, and provide the next state and reward after taking an action in a state. The discount factor scales the importance of future rewards, controlling how myopic the agent is. An agent makes decisions through a policy, a function which takes a state as input and outputs an action, possibly stochastically.

For many environments, the state space is too large for the policy to learn a map for each state. An example of this is the game of Go, one of the first breakthroughs of reinforcement learning, has a state space size of 10^{17} (Silver et al., 2016). To mitigate the issues caused by this, one commonly embeds the state space into a Euclidean space, commonly referred to as the agent's representation. If these representations are structured in a meaningful way, they can accelerate the agent's learning and planning. The field of representation learning focuses on learning these, and has experienced a surge of interest with recent large-scale endeavours.

In this thesis, we focus on different aspects of state similarity metrics, a way of quantifying the similarity between two states in MDPs. We study various metrics, and demonstrate that they capture their behavioural similarity through various theoretical results. We moreover show that these metrics can be used for effective representation learning, by using metric learning to represent the different metrics in the embedding space. We now provide in more detail an overview of the contributions.

1.1 Contributions

The MICo distance (Chapter 3)

Bisimulation metrics are a measure of the behavioural similarity of two states in an MDP, with rich theoretical properties. Unfortunately, they are prohibitively expensive to compute in practice unless one makes assumptions on the underlying environment. We propose a novel measure of state similarity known as the MICo distance, defined by a fixed-point definition akin to bisimulation metrics but designed to be efficiently computable. We relate the MICo distance to the bisimulation metric and prove similar theoretical properties. We then prove how it can be learnt efficiently from samples in settings with deterministic reward, and demonstrate that it can be learnt through a loss in function approximation settings.

Kernel similarity metrics (Chapter 4)

State similarity metrics such as bisimulation and the MICo distance are constructed by iterating a functional over a space of distances. In this chapter we take a different approach, and instead produce a kernel satisfying a fixed-point functional by iterating over the space of positive definite kernels. We then leverage reproducing kernel Hilbert space theory to recover a distance from this kernel, which we call the kernel similarity metric. We then prove an equivalence to a modification of the MICo distance, which allows us to expand upon the theory and provide alternate parametrisations for use in practice.

Distributional state similarity metrics (Chapter 5)

Distributional reinforcement learning is a recent framework in which one considers the entire distribution of returns, rather than simply the expected value. We adapt this perspective to state similarity metrics, and instead of considering the difference in expected rewards as done in bisimulation, MICo, and kernel similarity metrics, we consider the expected difference in reward distributions. This perspective allows us to produce novel connections to concepts in distributional reinforcement learning, such as distances between the return distributions at two states. We moreover extend the sampling results presented in Chapter 3, and prove that when the sampling procedure is performed in general environments, one attains these distributional metrics.

Metrics as value functions in auxiliary Markov decision processes (Chapter 6)

The connection between state similarity metrics and auxiliary Markov decision processes was first shown in (Ferns and Precup, 2014), which proved that the bisimulation metric between two states can be realized as the optimal value function in an auxiliary Markov decision process. This idea was further used in (Castro et al., 2021), which showed that the MICo distance corresponds to the policy value function in a given auxiliary Markov decision process. We formalize and generalize this framework, and show that many state similarity metrics can be constructed by a choice of a reward function and coupling. We provide a number of examples of these, including all state similarity metrics considered in this thesis as well as a number of novel distances. We then use this framework to prove a convergence result concerning convergence under changing policies (a setting which has not been considered in theory but is commonly applied in practice), and through the framework we show that this holds for all state similarity metrics considered.

Chapter 2

Background

In this chapter we review basic mathematical structures, then we review basic concepts of reinforcement learning and finally bisimulation and some of its variants.

2.1 Review of basic mathematical structures

2.1.1 Metric spaces

A metric space is a pair $(\mathcal{X}, d)$ of a set $\mathcal{X}$ with a function, called the *metric* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{\geq 0}$, such that for all $x, y, z \in \mathcal{X}$:

- $d(x, y) = 0 \iff x = y$ *Identity of indiscernibles*
- $d(x, y) = d(y, x)$ *Symmetry*
- $d(x, y) \leq d(x, z) + d(y, z)$ *Triangle inequality*

Various modifications of metrics can be formed by relaxing some of these constraints. *Pseudometric* spaces arise when the identity of indiscernibles is weakened to $x = y \implies d(x, y) = 0$, which corresponds to allowing distinct states to be at zero distance from each other. Almost all distance functions we are going to consider in this thesis will be pseudometrics, and so we will freely use the term “metric” when referring to pseudometrics.

2.1.2 Convergence and contraction

A metric d on a set $\mathcal{X}$ allows us to define topological notions on $\mathcal{X}$, such as the openness of sets, continuity, and convergence, which will be most important for our purposes. Given a sequence of points $(x_n)_{n\geq 0}$ in $\mathcal{X}$, we say that x_n converges to $x \in \mathcal{X}$, or $x_n \to x$, if for any $\varepsilon > 0$, there exists an $N > 0$ such that for all $n > N$, $d(x_n, x) < \varepsilon$. It is clear that the choice of metric d plays a critical role in the convergence properties of the space $(\mathcal{X}, d)$.

Every convergent sequence $(x_n)_{n\geq 0}$ is *Cauchy*, which means that for any $\varepsilon > 0$, there exists an $N > 0$ such that for all $n, m > N$, we have $d(x_n, x_m) < \varepsilon$. Intuitively, a Cauchy sequence is one in which points become arbitrarily close to one another, which one may expect should imply convergence. Indeed this does, however it may happen that the point to which it converges is not in $\mathcal{X}$, and so the sequence may not converge in $\mathcal{X}$. Indeed, consider the sequence $x_n = \frac{1}{n}$, where $\mathcal{X} = (0, 1]$. It is clear that x_n becomes arbitrarily close to 0, but $(x_n)_{n\geq 0}$ does not converge in $\mathcal{X}$ as $0 \notin \mathcal{X}$. With this in mind, one may desire that metric space $(\mathcal{X}, d)$ have the property that every Cauchy sequence converges; this can equivalently be seen as the space containing all of its limit points. If a metric space satisfies the property that every Cauchy sequence converges, it is called a *complete* metric space.

Complete metric spaces are also desirable as they are a setting in which we can use *Banach's fixed point theorem* (Banach, 1922). Let $(\mathcal{X}, d)$ be a complete metric space, and $T : \mathcal{X} \to \mathcal{X}$ be a function. A point $x \in \mathcal{X}$ is a fixed point of T if we have $T(x) = x$. We will see in later sections that whether fixed points exist, and finding them, is often of interest to us. The fixed point theorem provides a simple condition to verify this.

Definition 2.1.1. Let $(\mathcal{X}, d)$ be a metric space, and $T : \mathcal{X} \to \mathcal{X}$ be a function. T is said to be a contraction mapping with modulus $\beta \in (0, 1)$ if for all $x, y \in \mathcal{X}$, we have that

$$d(T(x), T(y)) \leq \beta d(x, y).$$

Theorem 2.1.2 (Banach fixed-point theorem). *Let $(\mathcal{X}, d)$ be a complete metric space, and $T : \mathcal{X} \to \mathcal{X}$ be a contraction mapping with modulus β. Then T admits a unique fixed point $x^* \in \mathcal{X}$. Moreover, for any $x_0 \in \mathcal{X}$, the sequence defined by $x_{n+1} = T(x_n)$ satisfies $x_n \to x^*$ as $n \to \infty$.*

2.1.3 Metrics on probability distributions

For this chapter and the rest of the thesis, we assume the reader is familiar with the basic concepts of measure-theoretic probability. Given a metric space $\mathcal{X}$ we write $\mathscr{P}(\mathcal{X})$ for the space of probability measures defined on the Borel sets of $\mathcal{X}$. Given two probability measures μ and ν on a set $\mathcal{X}$, a coupling $\lambda \in \mathscr{P}(\mathcal{X} \times \mathcal{X})$ of μ and ν is a joint distribution with marginals μ and ν. Formally, we have that for every measurable subset $\mathcal{A} \subset \mathcal{X}$,

$$\lambda(\mathcal{A} \times \mathcal{X}) = \mu(\mathcal{A}) \text{ and } \lambda(\mathcal{X} \times \mathcal{A}) = \nu(\mathcal{A}).$$

We define $\Lambda(\mu, \nu)$ to represent the set of all couplings of μ and ν. This set is always non-empty, in particular, the independent coupling $\lambda = \mu \times \nu$ always exists.

Couplings are essential for the definition of the Kantorovich metric $\mathcal{W}$ (Kantorovich and Rubinshtein, 1958) (also known as the Wasserstein metric), which lifts a metric d on $\mathcal{X}$ onto a metric on $\mathscr{P}(\mathcal{X})$. Given a metric d on $\mathcal{X}$, the Kantorovich metric is defined as

$$\mathcal{W}(d)(\mu, \nu) = \inf_{\lambda \in \Lambda(\mu, \nu)} \int d(x, y) \, \mathrm{d}\lambda(x, y).$$

The coupling which attains the infimum always exists, and is referred to as the *optimal coupling* of μ and ν (Villani, 2008).

Another distance on probability distributions which is defined through couplings is the total variation distance, defined for measures μ, ν as

$$TV(\mu, \nu) = \inf_{\lambda \in \Lambda(\mu, \nu)} \left\{ \mathbb{P}_{(X,Y) \sim \lambda} (X \neq Y) \right\}.$$

One important difference of the total variation distance and the Kantorovich distance is that the total variation does not depend on a base metric on $\mathcal{X}$, while the Kantorovich does. One way to phrase this is that for measures on metric spaces $(\mathcal{X}, d)$, the Kantorovich distance accounts for geometric properties of the space while the total variation does not. This can be seen through the following example.

Example 2.1.3. *Consider the metric space $(\mathbb{R}, |\cdot|)$. For $n \in \mathbb{N}$, let $\mu_n = \delta_{1/n}$, and $\mu = \delta_0$, where δ_x is the Dirac measure concentrated at $x \in \mathbb{R}$. Then intuitively, $(\mu_n)_{n\geq 0}$ is a sequence of Dirac measures approaching δ_0. This convergence is captured through the Kantorovich metric, as we have $\mathcal{W}(|\cdot|)(\mu_n, \mu) = \frac{1}{n}$, and we indeed have $\mu_n \to \mu$ in $\mathcal{W}(|\cdot|)$. However, we have that $TV(\mu_n, \mu) = 1$, and hence $\mu_n \not\to \mu$ in TV.*

An important class of metrics on probability distributions are known as *integral probability metrics* (Sriperumbudur et al., 2009b, 2012). They are parametrised through a choice of a real-valued function space $\mathscr{F}$, and for a given $\mathscr{F}$ the metric is given through

$$d_{\mathscr{F}}(\mu, \nu) = \sup_{f \in \mathscr{F}} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right|.$$

Both the Kantorovich and total variation distances are indeed integral probability metrics. The total variation distance corresponds to the choice $\mathscr{F} = \{f : \|f\|_\infty \leq 1\}$, which can be seen through some algebra and probabilistic arguments. The fact that the Kantorovich distance is an integral probability metric is less immediate. Given a metric space $(\mathcal{X}, d)$, let $\text{Lip}_1(d)$ be the set of functions which are 1-Lipschitz with respect to d, that is $\text{Lip}_1(d) = \{f : |f(x) - f(y)| \leq d(x, y) \ \forall x, y \in \mathcal{X}\}$. The fact that when $\mathscr{F} = \text{Lip}_1(d)$ corresponds to the Kantorovich metric is the celebrated Kantorovich-Rubenstein duality theorem (Kantorovich and Rubinshtein, 1958):

$$\inf_{\lambda \in \Lambda(\mu,\nu)} \int d(x, y) \, \mathsf{d}\lambda(x, y) = \sup_{f \in \text{Lip}_1(d)} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right|.$$

2.2 Bisimulation of labelled Markov processes

2.2.1 Labelled Markov processes

A stochastic process is a sequence of random variables, formally described as an indexed family

$$X_t : \Omega \to \mathcal{X},$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $\mathcal{X}$ is the state space, and $t \in T$ is the indexing set, commonly thought of as time. At each time t, the process induces a probability distribution P_t over $\mathcal{X}$, given by the distribution of the state at time t:

$$P_t(B) = \mathbb{P}\left(\{\omega : X_t(\omega) \in B\}\right).$$

This represents the probability that the process is in $B \subseteq \mathcal{X}$ at time t. One may also want to know the probability of being in a set at time t after observing the first $t-1$ values of the process, that is the conditional probability

$$P_t(B \,|\, x_0, \ldots, x_{t-1}) = \mathbb{P}\big(\{\omega : X_t(\omega) \in B\} \,\big|\, X_0 = x_0, \ldots, X_{t-1} = x_{t-1}\big).$$

If the process X_t has the *Markov property*, the previous expression can be simplified to only consider the previous timestep:

$$P_t(B \,|\, x_0, \ldots, x_{t-1}) = \mathbb{P}\big(\{\omega : X_t(\omega) \in B\} \,\big|\, X_{t-1} = x_{t-1}\big).$$

If the process X_t is *stationary* then the distribution does not change over time, meaning the conditional probability can be further simplified to

$$P_t(B \,|\, x_0, \ldots, x_{t-1}) = \mathbb{P}\big(\{\omega : X_1(\omega) \in B\} \,\big|\, X_0 = x_{t-1}\big).$$

In this setting, it is common to introduce the transition kernel $\mathcal{P} : \mathcal{X} \to \mathscr{P}(\mathbb{R})$, where $\mathcal{P}_x(B) \coloneqq \mathcal{P}(x)(B)$ is the probability of transitioning into B from state x. If $\mathcal{X}$ is finite, $\mathcal{P}$ can be represented by a $|\mathcal{X}| \times |\mathcal{X}|$ matrix.

A *labelled Markov process* is an augmentation of a stationary Markov process using a set of labels $\mathcal{A}$ (also known as actions). From a state x, one can choose a label $a \in \mathcal{A}$, which then determines the next-state dynamics. The previous transition kernel can be adapted to take labels into account, now taking the form $\mathcal{P} : \mathcal{A} \times \mathcal{X} \to \mathscr{P}(\mathbb{R})$, where $\mathcal{P}_x^a$ represents the distribution over states after choosing a in state x.

2.2.2 Bisimulation relations

Bisimulation was invented in the context of concurrency theory by Milner (Milner, 1980) and Park (Park, 1981). Probabilistic bisimulation (Larsen and Skou, 1991; Blute et al., 1997; Desharnais et al., 2002; Panangaden, 2009) (henceforth just called bisimulation) is an equivalence on the state space of a labelled Markov process, where two states are considered equivalent if the behaviour from the states are *indistinguishable*. To define indistinguishability, we demand that transition probabilities to equivalence classes should be the same for equivalent states; that is, the equivalence classes preserve the dynamics of the process. In addition if there are more observables, for example, rewards, those should match as well. Bisimulation for MDP's was defined in (Givan et al., 2003).

This intuition can now be transformed into a definition. We say that an equivalence relation R on $\mathcal{X}$ is a bisimulation relation if for any $x, y \in \mathcal{X}$, xRy implies that

$$\forall a \in \mathcal{A},\ \forall C \in \mathcal{X}/R,\ \mathcal{P}_x^a(C) = \mathcal{P}_y^a(C).$$

We say that states x and y are bisimilar if there exists a bisimulation relation R such that xRy. We remark that there exists at least one bisimulation relation, as the diagonal relation $\Delta = \{(x, x) : x \in \mathcal{X}\}$ is always a bisimulation relation, albeit the least interesting one. One refers to the largest bisimulation relation as $\sim$, which is often the one of interest.

This version is due to Larsen and Skou (Larsen and Skou, 1991) and the extension to continuous state spaces is due to (Blute et al., 1997; Desharnais et al., 2002).

2.3 Reinforcement Learning

2.3.1 Markov decision processes

A (discounted) Markov decision process (MDP) is the setting for most formalizations of reinforcement learning. An MDP is a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{P}_E, \gamma)$. The interaction of an agent in the MDP is as follows: at a given timestep t, the agent receives a state $X_t \in \mathcal{X}$, takes an action $A_t \in \mathcal{A}$, and then receives a reward $R_t \in \mathbb{R}$ and future state $X_{t+1} \in \mathcal{X}$ given the action A_t. The *environment dynamics* $\mathcal{P}_E : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R} \times \mathcal{X})$ dictates the transition dynamics of the environment, providing the joint distribution over rewards and states, given a state and action. The discount factor $\gamma \in [0, 1)$ specifies how much value is assigned to future rewards. Markov decision processes are labelled Markov processes augmented with a reward for taking an action in a state.

A simplifying assumption that is often made is that the reward and next state are conditionally independent given a state and action, meaning that

$$\mathbb{P}\left((R_t, X_{t+1}) = (r_t, x_{t+1}) \,\middle|\, x_t, a_t\right) = \mathbb{P}\left(R_t = r_t \,\middle|\, x_t, a_t\right) \mathbb{P}\left(X_{t+1} = x_{t+1} \,\middle|\, x_t, a_t\right);$$

equivalently, this means there exists a *reward probability* $\mathcal{P}_{\mathcal{R}} : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R})$ and *transition probability* $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$ such that the measure $\mathcal{P}_E$ decomposes into the product

$$\mathcal{P}_E = \mathcal{P}_{\mathcal{R}} \times \mathcal{P}.$$

A *policy* $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ is a mapping used by the agent to make decisions, and we write

$$A_t \sim \pi(\cdot \mid X_t)$$

to indicate that the action A_t is sampled from the probability distribution $\pi(X_t)$. We write $\Pi = \mathscr{P}(\mathcal{A})^{\mathcal{X}}$ to denote the space of all policies.

We will often make use of the random trajectory $(X_k, A_k, R_k)_{k\geq 0}$ when reasoning with MDPs. We will occasionally take expectations with respect to policies, written as $\mathbb{E}_\pi[f(X_t, A_t, R_t)]$, which should be read as the expectation, given that for all $k \geq 0$ we choose $A_k \sim \pi(\cdot|X_k)$, and receive $R_k \sim \mathcal{P}_\mathcal{R}(\cdot \,|\, X_k, A_k)$ and $X_{k+1} \sim \mathcal{P}(\cdot \,|\, X_k, A_k)$.

Other simplifying notation we will introduce is writing $\mathcal{R}^a_x$ and $\mathcal{R}^\pi_x$ as the random variable returns from a state from an action a or policy π. That is, one has $\mathcal{R}^a_x \sim \mathcal{P}_\mathcal{R}(\cdot \,|\, x, a)$, and $\mathcal{R}^\pi_x \sim \sum_{a\in\mathcal{A}} \pi(a \,|\, x)\mathcal{P}_\mathcal{R}(\cdot \,|\, x, a)$. We use r^a_x and r^π_x for the expectations of these random variables, that is $r^a_x = \mathbb{E}[\mathcal{R}^a_x]$ and $r^\pi_x = \mathbb{E}[\mathcal{R}^\pi_x]$. Lastly, we will write $\mathcal{P}^a_x \coloneqq \mathcal{P}(a)(x)$, and $\mathcal{P}^\pi_x = \sum_{a\in\mathcal{A}} \pi(a \,|\, x)\mathcal{P}^a_x$.

The *value* of a policy π is the expected total return an agent attains from following π, and is described by a function $V^\pi : \mathcal{X} \to \mathbb{R}$, such that for each $x \in \mathcal{X}$,

$$V^\pi(x) = \mathbb{E}_\pi\left[\sum_{t\geq 0} \gamma^t R_t \,\middle|\, X_0 = x\right],$$

A related quantity is the action-value function $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, which indicates the value of taking an action in a state, and then following the policy:

$$Q^\pi(x, a) = \mathbb{E}_\pi\left[\sum_{t\geq 0} \gamma^t R_t \,\middle|\, X_0 = x, A_0 = a\right].$$

A foundational relationship in reinforcement learning is the *Bellman equation*, which allows the value function of a state to be written recursively in terms of next states. It exists in two forms, for V^π and Q^π respectively:

$$V^\pi(x) = \mathbb{E}_\pi\left[R_0 + \gamma V^\pi(X_1) \,\middle|\, X_0 = x\right],$$
$$Q^\pi(x, a) = \mathbb{E}_\pi\left[R_0 + \gamma V^\pi(X_1) \,\middle|\, X_0 = x, A_0 = a\right].$$

Using the simplifying notation introduced earlier, we can write the Bellman equations as

$$V^\pi(x) = r_x^\pi + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^\pi} \left[V^\pi(x') \right],$$
$$Q^\pi(x,a) = r_x^a + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^a, a' \sim \pi(\cdot|x')} \left[Q^\pi(x', a') \right].$$

The Bellman optimality equations are obtained by taking a maximum of the above equations over all policies, and are given by

$$V^*(x) = \max_{\pi \in \Pi} V^\pi(x)$$
$$Q^*(x,a) = \max_{\pi \in \Pi} Q^\pi(x,a)$$

The Bellman operator T^π transforms the above equations into an operator over $\mathbb{R}^{\mathcal{X}}$ (or $\mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ - we will overload the use of T^π and let the type signature indicate which is being used), given by

$$T^\pi(V)(x) = r_x^\pi + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^\pi} \left[V(x') \right],$$
$$T^\pi(Q)(x,a) = r_x^a + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^a, a' \sim \pi(\cdot|x')} \left[Q(x', a') \right].$$

Written in this way, we see that Q^π and V^π are fixed points of T^π, and with some work one can also see that T^π is a contraction with modulus γ. As a corollary of Banach's fixed point theorem, one can choose V_0 arbitrarily and update $V_{k+1} = T^\pi V_k$, and converge to V^π, this is the algorithm known as *value iteration*.

2.3.2 MDP bisimulation

As Markov decision processes may be seen as Markov processes with rewards, bisimulation relations and metrics have a natural analogue in this setting.

Definition 2.3.1. An equivalence relation R on $\mathcal{X}$ is a *bisimulation relation* if

$$xRy \implies \forall a \in A, r_x^a = r_y^a \text{ and } \forall C \in \mathcal{X}/R,\, \mathcal{P}_x^a(C) = \mathcal{P}_y^a(C).$$

The stringency of bisimulation relations is apparent in the MDP case: if two states have bisimilar transition dynamics, that is we have that $\forall a$ and $\forall C \in \mathcal{X}/R$, $\mathcal{P}_x^a(C) = \mathcal{P}_y^a(C)$, but $|r_x^a - r_y^a| = \varepsilon > 0$, then x and y are in different equivalence classes. This motivates us to introduce a metric analogue of bisimulation. We will write $\mathcal{M}(\mathcal{X})$ to represent the space of pseudometrics on $\mathcal{X}$.

Definition 2.3.2. (Bisimulation metrics). For a given $c_R \in (0, \infty]$ and $c_T \in (0, 1)$, define $\mathcal{F} : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ as

$$\mathcal{F}(d)(x, y) = \max_{a \in \mathcal{A}} \left(c_R\, |r_x^a - r_y^a| + c_T\, \mathcal{W}(d)(\mathcal{P}_x^a, \mathcal{P}_y^a) \right).$$

Then $\mathcal{F}$ is a contraction in $\|\cdot\|_\infty$ with modulus c_T, and hence exhibits a unique fixed point $d^\sim$, which we denote a bisimulation metric. Justification for the term bisimulation metric follows from the fact that the kernel of $d^\sim$ is a bisimulation relation.

We will often take $c_R = 1$ and $c_T = \gamma$, as we will see that these correspond to value functions most naturally. We note that in the original definition presented in (Ferns et al., 2004) $\mathcal{M}(\mathcal{X})$ the constants c_R and c_T were restricted so that $c_R + c_T \leq 1$. This restriction was necessary for the proof of the existence of the metric, as the proof relied on the Knaster–Tarski fixed point theorem for lattices, and restricting $\mathcal{M}(\mathcal{X})$ to be 1-bounded made it a lattice. In this treatment however we rely on the Banach fixed point theorem, and do not require that elements of $\mathcal{M}(\mathcal{X})$ are bounded.

We will now discuss the term *behavioural metrics*. We say that d is a behavioural metric if $d(x, y)$ in some way measures the behavioural distance of x and y, where the behaviour of a state refers to a measure of the transition and reward dynamics from a state, rather than a naive distance which simply compares pixel values or other surface-level state differences. In the sense of bisimulation metrics, we now demonstrate that $d^\sim$ measures behavioural similarity in the sense of optimal value functions.

Theorem 2.3.3. *Suppose that $c_R \geq \gamma$, then we have that for any $x, y \in \mathcal{X}$,*

$$|V^*(x) - V^*(y)| \leq \frac{1}{1 - c_T} d^\sim(x, y).$$

Moreover, in the case that $c_T = \gamma$, we have

$$|V^*(x) - V^*(y)| \leq d^\sim(x, y).$$

The concept of a bisimulation metric was first introduced in (Desharnais et al., 1999, 2004) and adapted to MDP's in (Ferns et al., 2004)

2.3.3 On-policy bisimulation

Bisimulation considers equivalence across all possible actions, which is a strong notion of equivalence. In many settings, in a given state an agent may not be concerned with the behaviour under every possible action, but instead only with the actions which it may take under a given policy. On-policy bisimulation (Castro, 2020) was introduced to address this. The definition is a straightforward modification of MDP bisimulation, adapted to a given policy.

Definition 2.3.4 (On-policy bisimulation relations)**.** Let π be a fixed policy. An equivalence relation R on $\mathcal{X}$ is a *π-bisimulation relation* if

$$xRy \implies r_x^\pi = r_y^\pi \text{ and } \forall C \in \mathcal{X}/R, \mathcal{P}_x^\pi(C) = \mathcal{P}_y^\pi(C).$$

It is important to note that while the two definitions appear very similar, they have intrinsic differences. In particular, two states which are bisimilar need not be π-bisimilar, and two states which are π-bisimilar need not be bisimilar. To see this intuitively, consider two states which are bisimilar, then they are behaviourally identical under every action. But a given policy may select actions differently between the two states so that the expected rewards under the policy are different between the states, and in particular the two states are not π-bisimilar. On the other hand, two states may have different dynamics across different actions, and hence not be bisimilar, but the policy can balance the actions such that the states are π-bisimilar.

The π-bisimilarity equivalence relation, like ordinary bisimulation, is sensitive to small changes in the system parameters; so defining a metric in place of a relation is the natural next step.

Definition 2.3.5 (On-policy bisimulation metrics)**.** For a policy π on M, define $\mathcal{F}^\pi : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ as

$$\mathcal{F}^\pi(d)(x, y) = |r^\pi_x - r^\pi_y| + \gamma\mathcal{W}(d)(\mathcal{P}^\pi_x, \mathcal{P}^\pi_y).$$

Then $\mathcal{F}^\pi$ is a contraction with modulus γ, and hence admits a unique fixed point $d^\pi_\sim$, which we define to be the π-bisimulation metric.

It is straightforward to see that the kernel of $d^\pi_\sim$ is an on-policy bisimulation relation, justifying its name. Akin to bisimulation metrics, π-bisimulation metrics posess desirable continuity properties when it comes to *policy* value functions.

Proposition 2.3.6. *Let π be a policy, then for any $x, y \in \mathcal{X}$, we have that*

$$|V^\pi(x) - V^\pi(y)| \leq d^\pi_\sim(x, y).$$

2.3.4 Learning from samples

An important concept in reinforcement learning is that an agent often does not have access to the $\mathcal{P}$ and $\mathcal{R}$ beforehand, and instead must learn from a stream of samples $(X_k, A_k, R_k, X'_k)_{k\geq 0}$. As an example, computing the value function V^π through Bellman iteration, that is computing the sequence $V_{k+1} = T^\pi V_k$, requires knowledge of both $\mathcal{P}$ and $\mathcal{R}$ to compute the Bellman operator. Temporal difference learning is the sample-based counterpart to this operator, which builds a sequence of value function estimates $(V_k)_{k\geq 0}$ from a sequence of samples $(X_k, A_k, R_k, X_k')_{k\geq 0}$ as

$$V_{k+1}(X_k) = (1-\alpha_k)\,V_k + \alpha_k\,(R_k + \gamma V_k(X_k')),$$

and $V_{k+1}(x) = V_k(x)$ for $x \neq X_k$, where $(\alpha_k)_{k\geq 0}$ is a sequence of stepsizes satisfying the Robbins-Monro conditions.

In general, sample-based methods allow one to transform an operator that depends on knowledge of $\mathcal{P}$ and $\mathcal{R}$ into an incremental algorithm that depends only on samples. Formally, let $\mathcal{O}$ be an operator over some space of functions, whose fixed point F^* is of interest. Given a sample (X_k, A_k, R_k, X_k'), we construct a sample target $\hat{\mathcal{O}}$ which only depends on the sample. It is essential that the sample target is unbiased, in the sense that for any function F, we have

$$\mathbb{E}_\pi[\hat{\mathcal{O}}F] = \mathcal{O}\,F.$$

With this, one can construct a sequence of iterates $(F_k)_{k\geq 0}$ by choosing F_0 arbitrarily and setting

$$F_{k+1} = (1-\alpha_k)\,F_k + \alpha_k\,\hat{\mathcal{O}}(F_k),$$

where $(\alpha_k)_{k\geq 0}$ is a sequence of stepsizes. We then have convergence of $F_k \to F^*$ if the following conditions are met (Bertsekas and Tsitsiklis, 1996):

1. The stepsizes $(\alpha_k)_{k\geq 0}$ satisfy the Robbins-Monro conditions (Robbins and Monro, 1951), that is:

   $$\sum_{k\geq 1}\alpha_k = \infty,\ \sum_{k\geq 1}\alpha_k^2 < \infty,$$

   both with probability 1.

2. The noise process is a martingale difference sequence:

$$\mathbb{E}[\hat{\mathcal{O}}(F_k) - \mathcal{O}(F_k) \,\big|\, \alpha_0, \ldots, \alpha_k, F_0, \ldots, F_k] = 0.$$

3. The noise process has bounded conditional variance: there exists $A, B \in \mathbb{R}$ such that

$$\mathbb{E}[(\hat{\mathcal{O}}(F_k) - \mathcal{O}(F_k))^2 \,\big|\, \alpha_0, \ldots, \alpha_k, F_0, \ldots, F_k] \leq A + B\|F_k\|^2.$$

The process of learning from a stream of samples is also known as incremental learning or stochastic approximation, and has been vital in almost all modern reinforcement learning successes.

Chapter 3

MICo

3.1 Introduction

3.1.1 Drawbacks of the bisimulation metric

Using dynamic programming approaches, bisimulation metrics can be computed by iterating $d_{k+1} = \mathcal{F}(d_k)$ for classical bisimulation, or $d_{k+1} = \mathcal{F}^\pi(d_k)$ for on-policy bisimulation. These can compute an ε-approximation of the metric in in $O(|\mathcal{X}|^5|\mathcal{A}| \log \varepsilon / \log \gamma)$ time for classical bisimulation, and $O(|\mathcal{X}|^5 \log \varepsilon / \log \gamma)$ time for on-policy bisimulation (Castro et al., 2021). Being dynamic programming approaches however, these methods require exact knowledge of $\mathcal{P}$ and $\mathcal{R}$ for each iteration.

Various works have attempted to adapt learning bisimulation metrics in the online reinforcement-learning setting, however they either produce biased estimates of the metric (Ferns et al., 2006), or require assumptions such as transitions which are deterministic (Castro, 2020) or Gaussian (Zhang et al., 2021).

3.2 Definition

3.2.1 The Łukaszyk–Karmowski distance

In subsection 2.1.3, we discussed couplings, and noted that for any pair of probability distributions μ, ν on a set $\mathcal{X}$, the space of couplings $\Lambda(\mu, \nu)$ is never empty since the independent coupling $\lambda = \mu \times \nu$ always exists. We recall the definition of the Kantorovich metric which optimizes over the space of couplings when computing the distance:

$$\mathcal{W}(d)(\mu, \nu) = \inf_{\lambda \in \Lambda(\mu,\nu)} \int d(x, y) \, \mathsf{d}\lambda(x, y).$$

The main computational difficulty of calculating the Kantorovich distance comes from calculating this infimum, since for each pair of measures one must solve an optimization problem. The *Łukaszyk–Karmowski distance* d_{LK} (Łukaszyk, 2004) eschews this optimization, and instead considers the independent coupling between the measures. That is,

$$d_{LK}(d)(\mu, \nu) = \int d(x, y) \, \mathsf{d}(\mu \times \nu)(x, y),$$

or equivalently

$$d_{LK}(d)(\mu, \nu) = \mathop{\mathbb{E}}_{x \sim \mu, y \sim \nu}[d(x, y)].$$

Without the need for the optimization over couplings, the computation of d_{LK} reduces to the above computation, which is much more computationally efficient. When the base distance d is the Euclidean distance $\|\cdot\|$, the Łukaszyk–Karmowski distance has been used in econometrics, usually referred to as Gini's coefficient (Gini, 1912; Yitzhaki, 2003).

The computational advantage comes at a price, however. While the Kantorovich satisfies all the axioms of a proper metric, the Łukaszyk–Karmowski distance does not. In particular, it does not satisfy the identity of indiscernibles, but not in the same way as pseudometrics. In particular, one may find a measure μ such that $d_{LK}(d)(\mu, \mu) > 0$. One can in fact show that a measure μ satisfies $d_{LK}(d)(\mu, \mu) = 0$ if and only if μ is a Dirac measure, that is $\mu = \delta_x$ for some $x \in \mathcal{X}$ (the Dirac measure concentrated at $x \in \mathcal{X}$ is written as δ_x and assigns mass 1 to x, and mass 0 to the rest of $\mathcal{X}$). We will see that the other direction is true as well, and that $d_{LK}(d)(\mu, \mu)$ is in some sense a measure of *dispersion* of μ. One interpretation of this in the literature (Łukaszyk, 2004) is that d_{LK} captures the concept of *uncertainty*: given two random variables $X \sim \mu$, $Y \sim \nu$, unless μ and ν are point masses, observed values of X and Y are less likely to be equal depending on the dispersions of μ and ν. Hence d_{LK} captures a measure of uncertainty in the observed distance of X and Y, compared to a proper probability metric which would assign distance 0 if $Law(X) = Law(Y)$.

The concept of distance functions with non-zero self distances has been considered before, in particular through *partial metrics* (Matthews, 1994). A partial metric is a function $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ such that for any $x, y, z \in \mathcal{X}$:

- $0 \leq d(x, y)$ *Non-negativity*
- $d(x, x) \leq d(x, y)$ *Small self-distances*
- $d(x, y) = d(y, x)$ *Symmetry*
- if $d(x, x) = d(x, y) = d(y, y)$, then $x = y$ *Indistancy implies equality*
- $d(x, y) \leq d(x, z) + d(y, z) - d(z, z)$ *Modified triangle inequality*

We note that there were additional axioms added to this definition, rather than simply removing the requirement that $d(x, x) = 0$. This is due to the fact that this definition was constructed so that one can easily construct a proper metric $\tilde{d}$ from a partial metric d, given by $\tilde{d}(x, y) = d(x, y) - \frac{1}{2}(d(x, x) + d(y, y))$. We can now show that this definition is indeed too strong for the Łukaszyk–Karmowski distance, which we demonstrate in the following examples.

Example 3.2.1. *The Łukaszyk–Karmowski distance does not have small self-distances.*

Proof. Take $\mathcal{X} = [0, 1]$, $d = |\cdot|$, $\mu = \delta_{1/2}$, $\nu = U([0, 1])$. Then $d(\nu, \nu) = \frac{1}{3} > \frac{1}{4} = d(\mu, \nu)$. □

Example 3.2.2. *The Łukaszyk–Karmowski distance does not satisfy the modified triangle inequality.*

Proof. Take $\mathcal{X} = [0, 1]$, $d = |\cdot|$, $\mu = \delta_0$, $\nu = \delta_1$, $\eta = \frac{1}{2}(\delta_0 + \delta_1)$. Then we have

$$d(\mu, \nu) = 1 > \frac{1}{2} = d(\mu, \eta) + d(\nu, \eta) - d(\eta, \eta),$$

breaking the inequality. □

To account for this, Castro et al. (2021) introduced a new notion of distance known as *diffuse metrics*. A diffuse metric is a function $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ such that for any $x, y, z \in \mathcal{X}$:

- $0 \leq d(x, y)$ *Non-negativity*
- $d(x, y) = d(y, x)$ *Symmetry*
- $d(x, y) \leq d(x, z) + d(y, z)$ *Triangle inequality*

It is straightforward to see that the Łukaszyk–Karmowski distance is a diffuse metric. In addition, an attractive property of the Łukaszyk–Karmowski distance for the reinforcement learning setting is that it lends itself readily to stochastic approximation. Given a stream of samples $(x_n)_{n\geq 1}$, $(y_n)_{n\geq 1}$ from random variables $X \sim \mu$ and $Y \sim \nu$ respectively, a base metric d, and a sequence of step sizes $(\alpha_n)_{n\geq 1}$ satisfying the Robbins-Monro conditions, one can construct a sequence of iterates (d_n) defined by

$$d_n = (1 - \alpha_n)\, d_{n-1} + \alpha_n\, d(x_n, y_n),$$

with $d_0 = 0$. Then we have $d_n \to d_{LK}(d)(\mu, \nu)$ as $n \to \infty$ (Robbins and Monro, 1951).

3.2.2 The MICo distance

As the most expensive step of computing bisimulation distances comes from the computation of the Kantorovich, it is a natural next step to use the Łukaszyk–Karmowski distance to simplify the computation. This is the idea behind the MICo distance. Given a set $\mathcal{X}$, we will use the notation $\mathcal{M}^{diff}(\mathcal{X})$ to represent the space of diffuse metrics on $\mathcal{X}$. We can now use this to define the MICo distance.

Definition 3.2.3. Given a policy π, the MICo operator $T_M^\pi : \mathcal{M}^{diff}(\mathcal{X}) \to \mathcal{M}^{diff}(\mathcal{X})$ is given by

$$T_M^\pi(U)(x,y) = |r_x^\pi - r_y^\pi| + \gamma\, d_{LK}(U)(\mathcal{P}_x^\pi, \mathcal{P}_y^\pi)$$

It is straightforward to see that T_M^π maps $\mathcal{M}^{diff}(\mathcal{X})$ into $\mathcal{M}^{diff}(\mathcal{X})$. From the following proposition, one can apply Banach's fixed point theorem to see the existence of a unique fixed point U^π which satisfies

$$U^\pi(x,y) = |r_x^\pi - r_y^\pi| + \gamma\, d_{LK}(U^\pi)(\mathcal{P}_x^\pi, \mathcal{P}_y^\pi).$$

Proposition 3.2.4. *The MICo operator T_M^π is a contraction with modulus γ.*

Proof. For $U, U' \in \mathcal{M}^{diff}(\mathcal{X})$,

$$\begin{aligned}
\|T_M^\pi(U) - T_M^\pi(U')\|_\infty &= \sup_{x,y\in\mathcal{X}} |T_M^\pi(U)(x,y) - T_M^\pi(U')(x,y)| \\
&= \sup_{x,y\in\mathcal{X}} \left|\gamma d_{LK}(U)(\mathcal{P}_x^\pi, \mathcal{P}_y^\pi) - \gamma d_{LK}(U')(\mathcal{P}_x^\pi, \mathcal{P}_y^\pi)\right| \\
&\le \gamma \sup_{x,y\in\mathcal{X}} \left\{ \mathop{\mathbb{E}}_{x'\sim\mathcal{P}_x^\pi, y'\sim\mathcal{P}_y^\pi} \left[|U(x',y') - U'(x',y')|\right] \right\} \\
&\le \gamma \|U - U'\|_\infty.
\end{aligned}$$

□

3.2.3 Properties of the MICo distance

We recall that one of the appealing properties of bisimulation (both original and on-policy), was that it upper-bounded the absolute distance between value functions (optimal or on-policy). We now present a theorem that shows that this is indeed the case for the MICo distance as well.

Theorem 3.2.5 (Value function upper bound)**.** *The MICo distance upper bounds the absolute difference between policy-value functions. That is, for $x, y \in \mathcal{X}$, we have*

$$|V^\pi(x) - V^\pi(y)| \leq U^\pi(x, y).$$

Proof. Let $V_0 : \mathcal{X} \to \mathbb{R}$ be defined as $V_0(x) = 0$ for all $x \in \mathcal{X}$, and $U_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be defined as $U_0(x, y) = 0$ for all $x, y \in \mathcal{X}$. We can then define sequences $(V_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ by $V_{k+1} = T^\pi V_k$, $U_{k+1} = T^\pi_M(U_k)$. From Banach's fixed point theorem, we know that both $(V_k) \to V^\pi$ and $(U_k) \to U^\pi$ uniformly. Let $x, y \in \mathcal{X}$ be arbitrary, we will now prove by induction that for all k, we have

$$|V_k(x) - V_k(y)| \leq U_k(x, y).$$

The base case $k = 0$ is immediate, since both the left hand side and right hand side are 0. We can now let $k \geq 0$ and assume the induction hypothesis for k, and see that

$$\begin{aligned}
|V_{k+1}(x) - V_{k+1}(y)| &= |T^\pi V_k(x) - T^\pi V_k(y)| \\
&= \left| r^\pi_x + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}^\pi_x} [V_k(x')] - \left(r^\pi_y + \gamma \mathop{\mathbb{E}}_{y' \sim \mathcal{P}^\pi_y} [V_k(y')] \right) \right| \\
&\leq |r^\pi_x - r^\pi_y| + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}^\pi_x, y' \sim \mathcal{P}^\pi_y} [|V_k(x') - V_k(y')|] \\
&\leq |r^\pi_x - r^\pi_y| + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}^\pi_x, y' \sim \mathcal{P}^\pi_y} [U_k(x', y')] \\
&= |r^\pi_x - r^\pi_y| + \gamma \, d_{LK}(U_k)(\mathcal{P}^\pi_x, \mathcal{P}^\pi_y) \\
&= T^\pi_M \, U_k(x, y) \\
&= U_{k+1}(x, y),
\end{aligned}$$

as desired. Hence we have that $|V_k(x) - V_k(y)| \leq U_k(x, y)$ for any $x, y \in \mathcal{X}$. We can now take $k \to \infty$ on both sides, and since $(U_k)_{k\geq 0}$ and $(V_k)_{k\geq 0}$ both converge uniformly, we can conclude

$$|V^\pi(x) - V^\pi(y)| \leq U^\pi(x, y).$$

□

As discussed in subsection 3.1.1, one of the most important drawbacks of bisimulation metrics is the fact that it cannot be estimated in an online fashion through samples. The following theorem now shows that this is not the case for the MICo distance, in the setting that the reward from a state is independent of the chosen action.

Theorem 3.2.6 (Online approximation)**.** *Let $M = (\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ be a Markov decision process, and π be a policy on M such that the reward from a state under the policy is constant (e.g. if the policy and reward is deterministic, or if the reward under all actions are equal). Let $(X_k, A_k, R_k, X'_k)_{k\geq 0}$ and $(Y_k, A'_k, R'_k, Y'_k)_{k\geq 0}$ be two sequences of transitions in M following π. Moreover, let $(\alpha_k)_{k\geq 0}$ be a sequence of stepsizes satisfying the Robbins-Monro conditions. Let $(U_k)_{k\geq 0}$ be a sequence of estimates constructed by choosing U_0 arbitrarily, and setting*

$$\begin{aligned} U_{k+1}(X_k, Y_k) &= (1 - \alpha_k)\, U_k(X_k, Y_k) + \alpha_k\, \left(|R_k - R'_k| + \gamma U_k(X'_k, Y'_k)\right) \\ U_{k+1}(x, y) &= U_k(x, y) \text{ if } (x, y) \neq (X_k, Y_k). \end{aligned}$$

Then if each pair of states is sampled infinitely often, we have that $U_k \to U^\pi$ with probability 1.

Proof. We can add and subtract the MICo operator applied to U_k, $T^\pi_M(U_k)$, from the target of the update rule, giving us

$$U_{k+1}(X_k, Y_k) = (1 - \alpha_k)U_k(X_k, Y_k) + \alpha_k \left(T^\pi_M(U_k) + \underbrace{|R_k - R'_k| + \gamma U_t(X'_k, Y'_k) - T^\pi_M(U_k)}_{w_k} \right).$$

We proceed following Bertsekas and Tsitsiklis (1996), and must show that: (i) w_k is a martingale difference sequence, and (ii) w_k has bounded conditional variance.

To begin, let $\mathcal{F}_k = \sigma\left(U_0, \cdots, U_k, w_0, \cdots w_k, \alpha_0, \cdots, \alpha_k, X_0, \cdots, X_k, Y_0, \cdots, Y_k\right)$ be the sigma algebra representing all the information contained in the sampling process up to time k. We first show that w_k is a martingale difference process:

$$\begin{aligned}\mathbb{E}\left[w_k | \mathcal{F}_k\right] &= \mathbb{E}\left[|R_k - R'_k| + \gamma U_t(X'_k, Y'_k) - T^\pi_M(U_k) \,\middle|\, \mathcal{F}_k\right] \\ &= \mathbb{E}\left[|R_k - R'_k| + \gamma U_t(X'_k, Y'_k) - T^\pi_M(U_k) \,\middle|\, X_k, U_k\right] \\ &= 0,\end{aligned}$$

where we used the fact that since R_k and R'_k are both almost surely constant, $\mathbb{E}[|R_k - R'_k|] = |\,\mathbb{E}[R_k] - \mathbb{E}[R'_k]|$.

We now must bound the conditional variance of the noise term, such that $\mathbb{E}[w_k^2 | \mathcal{F}_k] \leq A + B\|U_k\|^2$. We can write out

$$\begin{aligned}\mathbb{E}\left[w_k^2 | \mathcal{F}_k\right] &= \mathbb{E}\left[\left(|R_k - R'_k| + \gamma U_k(X'_k, Y'_k) - T^\pi_M(U_k)\right)^2 \,\middle|\, \mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\left(|R_k - R'_k| + \gamma U_k(X'_k, Y'_k)\right)^2 + \left(T^\pi_M(U_k)\right)^2 \,\middle|\, X_k, U_k\right] \\ &= \mathbb{E}\left[\left(|R_k - R'_k| + \gamma U_k(X'_k, Y'_k)\right)^2 \,\middle|\, X_k, U_k\right] + \mathbb{E}\left[\left(T^\pi_M(U_k)\right)^2 \,\middle|\, X_k, U_k\right] \\ &\leq A + B\|U_k\|^2.\end{aligned}$$

□

3.2.4 On self-distances

An interesting point to note is studying when a state has nonzero self-distance. We can write out the self-distance for a state x as

$$U^\pi(x, x) = \gamma\, d_{LK}(\mathcal{P}^\pi_x, \mathcal{P}^\pi_x),$$

so we see that the magnitude of the self-distance comes entirely from the Łukaszyk–Karmowski distance between the transition distributions. We know that the Łukaszyk–Karmowski self-distance measures the dispersion of a distribution, in the sense that it is minimised for a point mass, and maximised for a distribution which is maximally "spread out". Since we are measuring the self-distance of the transition distribution $\mathcal{P}_x^\pi$, we can see that $U^\pi(x, x)$ in a sense measures the dispersion of the transition dynamics from x.

As a converse to the previous point, we may inquire what happens when there is no dispersion in the transition distributions, that is when we have an MDP with deterministic transitions. In this case, we can see that the MICo distance U^π is equal to the on-policy bisimulation distance d^π; this is because of the fact that

$$\mathcal{W}(d)(\delta_x, \delta_y) = d_{LK}(d)(\delta_x, \delta_y) = d(x, y).$$

3.3 Representation learning using diffuse metrics

In many reinforcement learning applications, it is infeasible to compute value functions and related quantities in a table parametrised by states, which we will refer to as tabular reinforcement learning. Instead, one commonly uses a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ to simplify the setting and improve generalization (Lyle et al., 2021). It is common to refer to the matrix $\Phi = [\phi(x)]_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X} \times d}$ as the agent's *representation* (Tu and Recht, 2018).

For example, in the linear function approximation setting value functions are parametrised through a representation Φ and a weight vector w, so that for each $x \in \mathcal{X}$

$$V(x) = \langle \phi(x), w \rangle,$$

or concisely

$$V = \Phi w.$$

The field of *representation learning* is concerned with learning a good representation, allowing more efficient learning and generalization (Lan et al., 2022; Dabney et al., 2020). A behavioural distance perspective on representation learning could be that for a behavioural distance d, one should strive for the representation distance of states to approximate their behavioural distance, that is $\|\phi(x) - \phi(y)\| \approx d(x, y)$.

For the MICo distance, and more generally any behavioural diffuse metric, the above goal presents an issue: a given state may have positive self-distance, that is $d(x, x) > 0$, however for any representation ϕ, we have that $|\phi(x) - \phi(x)| = 0$. In other words, any learnt representation ϕ to approximate d will be biased. One may consider only learning d for pairs of distinct points, and not for self distances. This corresponds to learning $|\phi(x) - \phi(x)| \approx \tilde{d}(x, y)$, where

$$\tilde{d}(x, y) = \begin{cases} d(x, y) \text{ if } x \neq y \\ 0 \text{ if } x = y \end{cases}$$

While this may be an attractive idea, it is relatively naive. For example $\tilde{d}$ may not preserve continuity, as it can be seen as 'slicing off the diagonal'. A less crude way of approaching this issue can be done by subtracting the self-distances at each point, as done in Matthews (1994) and Cuturi (2013). This can be seen as a projection onto the space of functions with zero self-distances, and for a diffuse metric d, we define the projected reduced metric Πd as

$$\Pi d(x, y) = d(x, y) - \frac{1}{2}(d(x, x) + d(y, y)).$$

It is straightforward that Πd will satisfy both symmetry and zero self-distances. However, Πd may not satisfy the triangle inequality, depending on the properties of d. If d is a partial metric, then Πd will indeed satisfy the triangle inequality and be a proper metric. As the MICo distance may not be a partial metric (depending on the MDP), the projected MICo distance ΠU^π may not be a proper metric.

We now see that ΠU^π loses another desirable property of the MICo distance, the value function upper bound. This upper bound is important for a number of reasons: when it holds for a state metric d, it indicates that the value function is well-behaved (continuous) with respect to d. Moreover, as the goal of value-based reinforcement learning is learning the value function, having the upper bound hold suggests that agglomerating states with close distances under d does not hurt the ability to learn the value function.

Proposition 3.3.1. *The projected MICo distance ΠU^π may not satisfy the value function upper bound.*

Proof. Consider the two state MDP where $\mathcal{X} = \{x, y\}$, $\mathcal{A} = \{a\}$, $\mathcal{P}_x^a = \frac{1}{2}(\delta_x + \delta_y)$, $\mathcal{P}_y^a = \delta_y$, and both $\mathcal{R}_x^a = 1$, $\mathcal{R}_y^a = 0$ almost surely. Since there is a single action, there exists a single policy π. One can calculate

$$V^\pi(x) = \frac{1}{1-\frac{\gamma}{2}},\ V^\pi(y) = 0,\ U^\pi(x,y) = \frac{1}{1-\frac{\gamma}{2}},\ U^\pi(x,x) = \frac{\gamma}{2(1-\frac{\gamma}{4})(1-\frac{\gamma}{2})},\ U^\pi(y,y) = 0.$$

This gives us that

$$|V^\pi(x) - V^\pi(y)| = \frac{1}{1-\frac{\gamma}{2}} > \frac{1}{1-\frac{\gamma}{4}} = \Pi U^\pi(x,y).$$

□

Despite this negative result, one can inquire how much this bound is broken in practice, and to what extent. We study this, by analyzing how much the bound is broken on average, across a class of random MDPs. The choice of random MDP used are Garnet MDPs (Archibald et al., 1995; Piot et al., 2014), which are built through the following construction:

1. Fix a number of states $n_\mathcal{X}$ and a number of actions $n_\mathcal{A}$
2. For each $(x,a) \in \{1, \dots n_\mathcal{X}\} \times \{1, \dots, n_\mathcal{A}\}$, choose a branching factor $b_{x,a}$ uniformly in $\{1, \dots, n_\mathcal{X}\}$
3. For each $(x,a) \in \{1, \dots n_\mathcal{X}\} \times \{1, \dots, n_\mathcal{A}\}$, construct $\mathcal{P}_x^a$ by uniformly choosing $b_{x,a}$ states with replacement and creating a probability distribution from these samples
4. For each $(x,a) \in \{1, \dots n_\mathcal{X}\} \times \{1, \dots, n_\mathcal{A}\}$, choose r_x^a uniformly from $U([0,1])$.

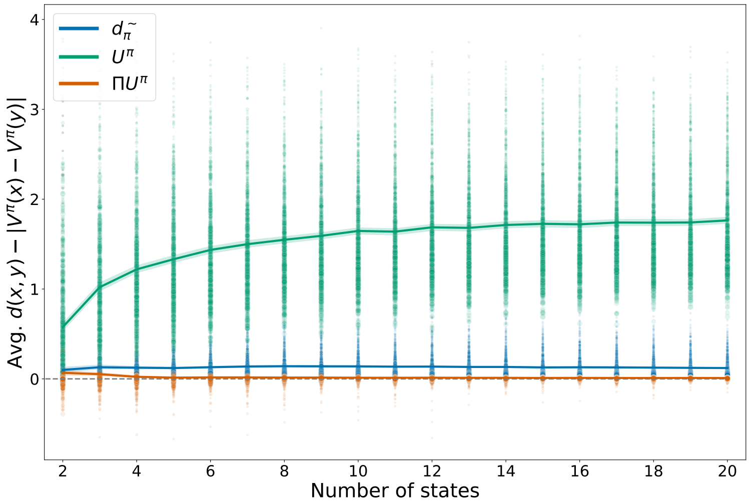


Figure 3.1: The gap between the difference in values and the various distances for Garnet MDPs with varying numbers of states and actions.

We vary the number of states $n_\mathcal{X}$ from 2 to 20, and for each MDP sample 100 stochastic policies $(\pi_k)_{k=1}^{100}$. For each Garnet MDP we compute the average upper bound gap across the 100 policies, computed as

$$\frac{1}{100|n_\mathcal{X}|^2}\sum_{k=1}^{100}\sum_{i=1}^{n_\mathcal{X}}\sum_{j=1}^{n_\mathcal{X}}\left(d(x_i,x_j) - |V^\pi(x_i) - V^\pi(x_j)|\right),$$

where d is a distance function (we consider U^π, ΠU^π, and $d_\pi^\sim$). We note that we are measuring the signed difference: if it is negative it indicates that the bound is being broken, and the magnitude of the value indicates how vacuous the bound is. We plot the result of this in Figure 3.1. As can be seen in the graph, although ΠU^π is the only distance which does not have the value function upper bound property, it is the closest approximate to the absolute value function difference, which is an indication that it produces effective features.

3.4 Applications to deep reinforcement learning

3.4.1 Deep Q Networks

Deep Q learning is an extension of the Q-learning algorithm to leverage the power of neural networks. The original architecture is DQN (Mnih et al., 2015), which represents the Q function as a neural network made up of five hidden layers: three convolutional layers, and two dense layers. The final dense layer has width $|\mathcal{A}|$, giving one output per action. The network can be written as a composition of two functions, a representation function ϕ_ω which represents the three convolutional layers with parameters ω, and the function approximator ψ_ξ which represents the two dense layers with parameters ξ. For an input state x, the network outputs $Q_{\omega,\xi}(x,\cdot) = \psi_\xi(\phi_\omega(x))$, where we use the subscripts ω and ξ to indicate that the function depends on both sets of parameters.

As the agent interacts in the environment, transition tuples of the form (x, a, r, x') are stored in a replay buffer $\mathcal{D}$. The learning process learns Q^* using Bellman's optimality equation

$$Q^*(x,a) = r_x^a + \gamma \max_{a' \in \mathcal{A}} \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^a} [Q^*(x', a')],$$

by using the fact that if $Q_{\omega,\xi}$ can be learnt so that

$$Q_{\omega,\xi}(x,a) \approx r_x^a + \gamma \max_{a' \in \mathcal{A}} \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^a} [Q_{\omega,\xi}(x', a')],$$

then $Q_{\omega,\xi} \approx Q^*$. To learn this, the network minimizes the following loss across the replay buffer:

$$\mathcal{L}(\omega,\xi) = \mathop{\mathbb{E}}_{(x,a,r,x') \sim \mathcal{D}} \left[\left(Q_{\xi,\omega}(x,a) - (r + \max_{a' \in \mathcal{A}} Q_{\xi,\omega}(x', a')) \right)^2 \right].$$

The value $r + \max_{a' \in \mathcal{A}} Q_{\xi,\omega}(x', a')$ is known as the target, and can be seen as the learning objective of the network. The fact that the objective depends on the current network itself is a phenomenon known as *bootstrapping* (Sutton and Barto, 2018). Since the target is changing with the current value, this can lead to unstable training. One way that Mnih et al. (2015) accounted for this was by using *target networks*: every 8000 evironment steps, they 'freeze' a snapshot of the parameters ω and ξ and refer to these frozen parameters as $\bar{\omega}$ and $\bar{\xi}$. These frozen parameters are then used in the target, and the loss becomes

$$\mathcal{L}(\omega, \xi) = \mathop{\mathbb{E}}_{(x,a,r,x')\sim\mathcal{D}} \left[\left(Q_{\xi,\omega}(x,a) - (r + \max_{a' \in \mathcal{A}} Q_{\bar{\xi},\bar{\omega}}(x',a')) \right)^2 \right].$$

3.4.2 The MICo loss

Recalling the goal of representation learning, our goal for the MICo loss would be to adapt the learned representations so that $d(\phi_\omega(x), \phi_\omega(y)) \approx \Pi U^\pi(x,y)$. We will see that choosing d to be the cosine distance between the embeddings lends itself nicely to a neural network parametrisation, where the cosine distance $\theta(x,y)$ is the angle between the vectors x and $y \in \mathbb{R}^d$, given by

$$\theta(x,y) = \arccos\left(\frac{x \cdot y}{\|x\|\|y\|}\right).$$

One difficulty that arises is that to produce an unbiased esimate of $\Pi U^\pi(x,y)$, we require at least two samples from x and y. This is because $\Pi U^\pi(x,y)$ contains the self-distance terms, $U^\pi(x,x)$ and $U^\pi(y,y)$. Looking at $U^\pi(x,x)$, we can write out

$$U^\pi(x,x) = d_{LK}(U^\pi)(\mathcal{P}^\pi_x, \mathcal{P}^\pi_x)$$

is the Łukaszyk–Karmowski distance between the transition distribution from x and itself. This is impossible to estimate unbiasedly with only a single sample of $\mathcal{P}^\pi_x$, since the estimate from a single sample will always be 0. In deep reinforcement learning, it is extremely rare to visit a state twice, and so expecting to learn ΠU^π directly is unrealistic.

As a result of this, we will need to learn ΠU^π implicitly. To do this, we will learn U^π, and parametrise it in the neural network as

$$U_\omega(x,y) = \frac{\|\phi_\omega(x)\| + \|\phi_\omega(y)\|}{2} + \beta\, \theta(\phi_\omega(x), \phi_\omega(y)).$$

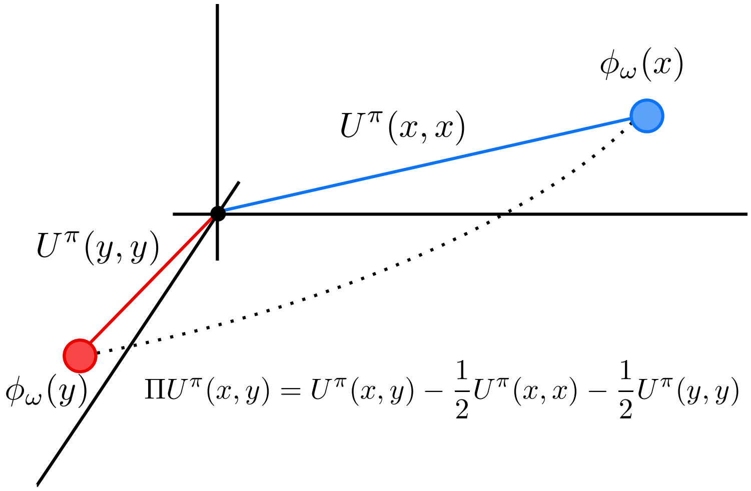

$\phi_\omega(x)$
$U^\pi(x,x)$
$U^\pi(y,y)$
$\phi_\omega(y)$
$\Pi U^\pi(x,y) = U^\pi(x,y) - \frac{1}{2}U^\pi(x,x) - \frac{1}{2}U^\pi(y,y)$

Figure 3.2: A visualization of the reduced MICo distance as the angular distance between representations.

We note that we only use the parameters ω here, the parametrisation of U^π lives entirely in the representation component of the neural network. By parametrising U_ω in this way, we can see that

$$\begin{aligned}\Pi U_\omega(x,y) &= U_\omega(x,y) - \frac{1}{2}(U_\omega(x,x) + U_\omega(y,y)) \\ &= \beta\,\theta(\phi_\omega(x), \phi_\omega(y)),\end{aligned}$$

so that we are indeed learning angular distances in ϕ which approximate the reduced MICo distance. β is a scalar hyperparameter which represents how much weight should be applied to the angular distance term. We present a diagram visualizing how ΠU^π models the angular distance in Figure 3.2.

With this parametrisation, we can now define the learning process with which U_ω is learnt. The procedure follows the outline of how the Bellman equation was transformed into a loss in subsection 3.4.1. We know that U^π satisfies the recursive equation

$$U^\pi(x,y) = |r^\pi_x + r^\pi_y| + \gamma \underset{x' \sim \mathcal{P}^\pi_x, y' \sim \mathcal{P}^\pi_y}{\mathbb{E}} [U^\pi(x', y')].$$

We can adapt this into a loss for deep reinforcement learning, letting $\bar{\omega}$ be the frozen parameters for ω, our learning target for pair of transitions $\langle x, r_x, x' \rangle$, $\langle y, r_y, y' \rangle$ is

$$|r_x - r_y| + \gamma U_{\bar{\omega}}(x', y').$$

Using this, we can construct the metric loss $\mathcal{L}_{\text{MICo}}$ as

$$\mathcal{L}_{\text{MICo}}(\omega) = \mathop{\mathbb{E}}_{\langle x, r_x, x' \rangle, \langle y, r_y, y' \rangle \sim \mathcal{D}} \left[\Big(U_\omega(x, y) - \big(|r_x - r_y| + \gamma U_{\bar{\omega}}(x', y') \big) \Big)^2 \right].$$

We refer to the loss used by an agent to train its behaviour as $\mathcal{L}_{\text{TD}}$, the *temporal difference* loss. For example, $\mathcal{L}_{\text{TD}}$ for DQN is the loss we derived at the end of subsection 3.4.1. To use the MICo loss for an agent, we combine the losses using a parameter α:

$$\mathcal{L}_{\text{Total}}(\xi, \omega) = (1 - \alpha)\, \mathcal{L}_{\text{TD}}(\xi, \omega) + \alpha\, \mathcal{L}_{\text{MICo}}(\omega).$$

There are two important things to note regarding the construction of this total loss. Firstly, $\mathcal{L}_{\text{MICo}}$ only depends on ω, so it directly shapes the representations learnt without being affected by the function approximator's parameters ξ. Secondly, the construction of the MICo loss does not depend on the choice of agent used, and so it can be applied to any existing agent.

3.4.3 Empirical performance

To study the empirical performance of the MICo loss, we add it as an auxiliary loss (as shown in the equation for $\mathcal{L}_{\text{Total}}$) to various value-based agents, and evaluate across the Atari benchmark (Bellemare et al., 2013; Machado et al., 2018). We add the loss to all agents provided in the Dopamine (Castro et al., 2018) library, which are: DQN (Mnih et al., 2015), Rainbow (Hessel et al., 2018), QR-DQN (Dabney et al., 2018b), IQN (Dabney et al., 2018a), and M-IQN (Vieillard et al., 2020). We report the score of the agent augmented with the MICo loss compared to the original agent across all 60 games and all 5 agents in Figure 3.3. As seen in the figure, the performance of all 5 agents were improved on average, including M-IQN, which is the state-of-the-art value-based agent on Atari at the time of this writing.

3.5 Discussion and future work

In this chapter, we introduced a novel distance for computing state similarity in Markov decision processes, importantly one that is computable through stochastic approximation and hence applicable in large-scale settings. We describe a process to transform the distance into a loss, which can be applied to any agent, and we present results using the loss across various agents on the Atari benchmark.

There are a number of directions which future work can be taken. Firstly, we recall that in deep reinforcement learning settings, the reduced distance ΠU^π cannot be learnt directly due to the fact that we almost never visit the same state twice. Although we are able to avoid this issue by learning ΠU^π implicitly, this is not an entirely satisfactory solution. One possible way around this is to leverage model-based reinforcement learning, in particular recent advances such as Dreamer (Hafner et al., 2019) seem promising, since we would not have to 'naturally' visit a state twice, we can instead sample two trajectories from a state to estimate self-distances.

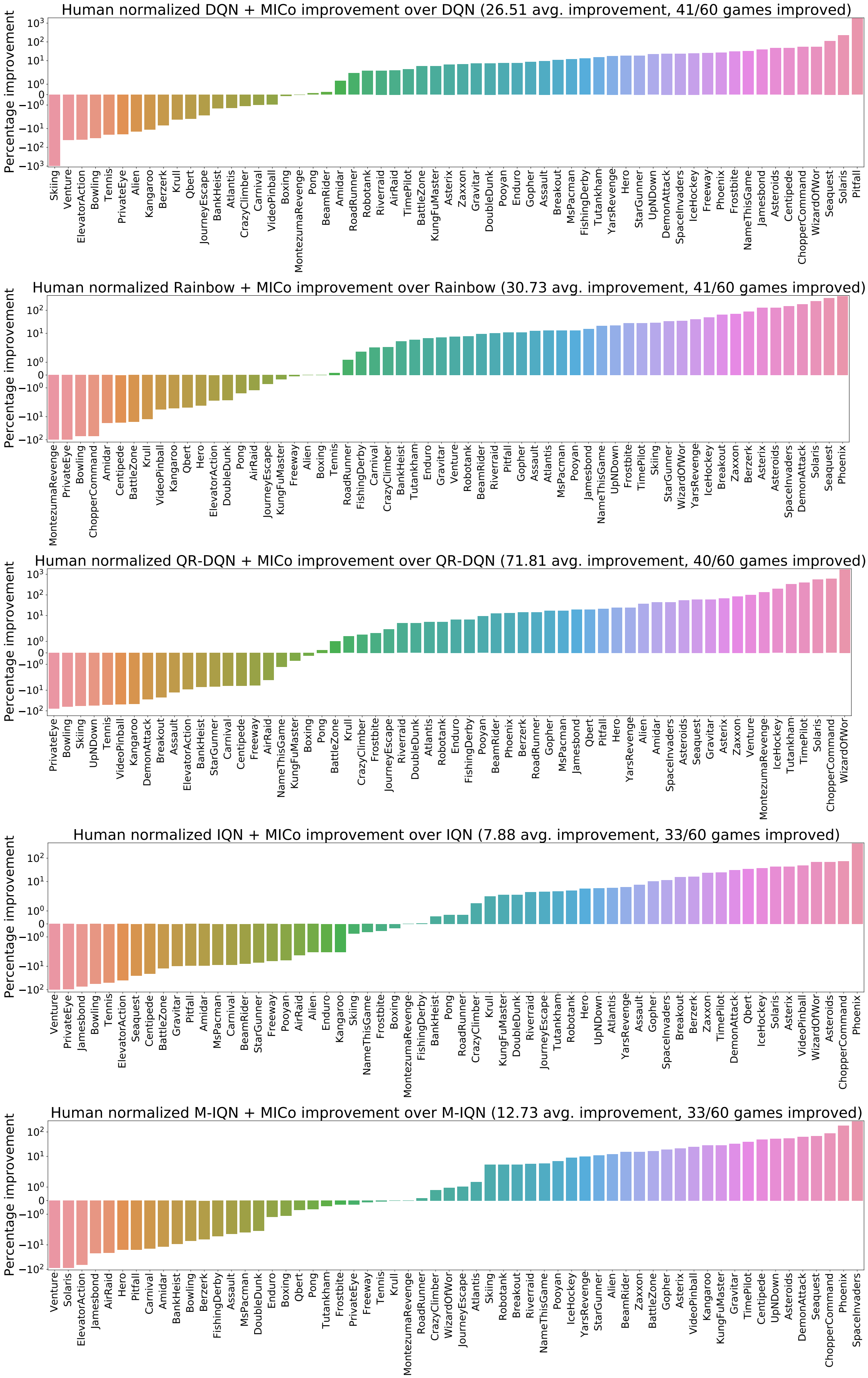
A second direction is a study of the looseness of the value function upper bound, inspired by Figure 3.1. In previous work on state similarity metrics (including this one), the primary relationship of interest is whether a distance d upper bounds the value function, that is

$$|V^\pi(x) - V^\pi(y)| \leq d(x, y).$$

However, as suggested from Figure 3.1, whether this upper bound holds may not be the most important question. It may be more important whether we have that

$$\Big| |V^\pi(x) - V^\pi(y)| - d(x, y) \Big|$$

is small. For an extreme example of the upper bound alone not being sufficient, consider the setting where the reward is bounded in $[0, 1]$. Then $V^\pi \in [0, \frac{1}{1-\gamma}]$, and so the distance $d(x, y) = \frac{2}{1-\gamma}\mathbb{1}_{x \neq y}$ upper bounds $|V^\pi(x) - V^\pi(y)|$, but is an uninteresting and trivial similarity metric.


Human normalized DQN + MICo improvement over DQN (26.51 avg. improvement, 41/60 games improved)

Human normalized Rainbow + MICo improvement over Rainbow (30.73 avg. improvement, 41/60 games improved)

Human normalized QR-DQN + MICo improvement over QR-DQN (71.81 avg. improvement, 40/60 games improved)

Human normalized IQN + MICo improvement over IQN (7.88 avg. improvement, 33/60 games improved)

Human normalized M-IQN + MICo improvement over M-IQN (12.73 avg. improvement, 33/60 games improved)

Figure 3.3: Percentage improvement in returns when adding $\mathcal{L}_{\text{MICo}}$ to various agents where the results are averaged over 5 independent runs.

Chapter 4

A Kernel Perspective

In this chapter we consider a slightly modified objective, and learn similarity measures on MDPs, rather than distance measures. We accomplish this using theory of positive-definite kernels, and prove that this approach allows us to recover a distance function from a Hilbert space embedding. We then prove equivalence of this distance function to the reduced MICo distance ΠU^{π} from Chapter 3, and discuss possible directions this equivalence gives us.

4.1 Background

In this section we review mathematical background covering vector spaces, reproducing kernel Hilbert spaces, the MMD, and its equivalence to the energy distance.

4.1.1 Hilbert spaces

A (real) normed space is a vector space V with a function $\|\cdot\| : V \to \mathbb{R}$, which satisfies the following for all $x, y \in V$, $\alpha \in \mathbb{R}$:

- $\|x\| \geq 0$ *Positivity*
- $\|x\| = 0 \iff x = 0$ *Identity of indiscernibles*

- $\|\alpha x\| = |\alpha|\|x\|$ *Absolute homogeneity with respect to scalar multiplication*

- $\|x + y\| \leq \|x\| + \|y\|$ *Triangle inequality*

A normed space is a stronger notion than a metric space, since a norm induces a metric d through $d(x, y) = \|x - y\|$. An inner product space is a stronger notion of a normed space, which is described as a vector space V with a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ such that for all $x, y, z \in V$, $\alpha, \beta \in \mathbb{R}$:

- $\langle x, y \rangle = \langle y, x \rangle$ *Symmetry*

- $\langle x, \alpha y + \beta z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$ *Linearity in the first argument*

- $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$ *Positive definiteness*

An inner product induces a normed space through $\|x\| = \langle x, x \rangle^{1/2}$. If an inner product space induces a normed space whose topology is complete, then the space $(V, \langle \cdot, \cdot \rangle)$ is referred to as a *Hilbert space*. A normed space whose topology is complete is referred to as a *Banach space*, and we remark that Hilbert spaces are a proper subset of Banach spaces.

Hilbert spaces have many desirable properties, one which will become important for the following theory is the *Riesz representation theorem*. Given a Hilbert space V, a map $T : V \to \mathbb{R}$ is linear if:

$$T(\alpha x + \beta y) = \alpha\, T(x) + \beta\, T(y) \text{ for all } x, y \in V,\ \alpha, \beta \in \mathbb{R}.$$

Continuity is easy to verify for linear maps: a linear map T is continuous if and only if T is bounded, meaning that there exists $C \in \mathbb{R}$ such that

$$\|T(x)\| \leq C\|x\|, \text{ for all } x \in V.$$

The set of all continuous linear operators on V is known as the dual space of V, and often referred to as $V^\star$. The Riesz representation theorem states that if V is a Hilbert space, then V and $V^\star$ are isometrically isomorphic. Equivalently, this means that for any linear operator $T : V \to \mathbb{R}$, there exists a unique $x_T \in V$ such that

$$\langle x, x_T \rangle = T(x) \text{ for all } x \in V.$$

4.1.2 Reproducing kernel Hilbert spaces and the MMD

Let $\mathcal{X}$ be an arbitary set and $\mathcal{H}$ be a Hilbert space of real functions on $\mathcal{X}$. For a point $x \in \mathcal{X}$, the evaluation functional $L_x : \mathcal{H} \to \mathbb{R}$ is defined by

$$L_x(f) = f(x).$$

If L_x is a continuous functional for all $x \in \mathcal{X}$, we say that $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) (Schölkopf et al., 2018; Aronszajn, 1950). Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space, then for each $x \in \mathcal{X}$, L_x is linear and continuous, and the Riesz representation theorem implies that there exists a unique $k_x \in \mathcal{H}$ such that

$$L_x(f) = \langle f, k_x \rangle_{\mathcal{H}}.$$

Since $k_x \in \mathcal{H}$, we can write

$$k_x(y) = L_y(k_x) = \langle k_x, k_y \rangle_{\mathcal{H}}.$$

This is used to define the *reproducing kernel* k of $\mathcal{H}$ as $k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}}$. We will sometimes write $k_{\mathcal{H}}$ to emphasize the dependence of the kernel on the Hilbert space. One can note that the functions k_x and k_y above can be recovered as the kernel fixed at a single point, that is $k_x = k(x, \cdot) \in \mathcal{H}$, and $k_y = k(y, \cdot) \in \mathcal{H}$. This is where the *reproducing property* comes from, as we see that k 'reproduces' itself:

$$k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}}.$$

In the previous paragraphs, we began with a Hilbert space of functions whose evaluation functional was continuous and obtained a reproducing kernel for this space. Going in the opposite direction is also possible; that is beginning with a positive definite kernel on a set and constructing a Hilbert space of functions, this is known as the *Moore-Aronszajn theorem* (Aronszajn, 1950). To begin, we define a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be a positive definite kernel if it is symmetric and positive definite[1]: for any $\{x_1, \ldots, x_n\} \in \mathcal{X}$, $\{c_1, \ldots, c_n\} \in \mathbb{R}$, we have that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

We will often use *kernel* as a shorthand for positive definite kernel. Given a kernel k on $\mathcal{X}$, we can construct a RKHS of functions $\mathcal{H}_k$ through the following steps:

(i) Construct a vector space of real-valued functions on $\mathcal{X}$ of the form $\{k(x, \cdot) : x \in \mathcal{X}\}$

(ii) Equip this space with an inner product given by $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_k} = k(x, y)$

(iii) Take the completion of the vector space with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$

The Hilbert space we obtain at the end of step (iii) is the reproducing kernel Hilbert space for k.

It is common to introduce the notation $\varphi(x) := k(x, \cdot)$, where $\varphi : \mathcal{X} \to \mathcal{H}$ is often called the *feature map*, and $\varphi(x)$ is understood as the *embedding* of x in $\mathcal{H}$. One can also embed probability distributions on $\mathcal{X}$ in $\mathcal{H}$. Given a probability distribution μ on $\mathcal{X}$, one can define the embedding of μ, $\Phi(\mu) \in \mathcal{H}$ as

$$\Phi(\mu) = \mathop{\mathbb{E}}_{X \sim \mu}[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) d\mu(x),$$

[1] We remark that the definition of positive definite is not consistent across the literature. We follow the convention of the kernel methods community, and define a function to be strictly positive definite if the inequality is strict unless $c_1 = \cdots = c_n = 0$. In the linear algebra and optimization communities however, this is referred to as positive definite, and the definition provided is referred to as positive semidefinite.

where the integral taken is a Bochner integral[2] , as we are integrating over $\mathcal{H}$-valued functions. The embeddings of measures into $\mathcal{H}$ allow one to easily compute integrals, as one can show using the Riesz representation theorem that for $f \in \mathcal{H}$, one has

$$\int_{\mathcal{X}} f d\mu = \langle f, \Phi(\mu) \rangle_{\mathcal{H}_k}.$$

These embeddings also allow us to define metrics on $\mathcal{X}$ and $\mathscr{P}(\mathcal{X})$ by looking at the Hilbert space distance of their embeddings. We will denote the distance on $\mathcal{X}$ that this induces as ρ, so that we have

$$\rho_k(x, y) = \|\varphi(x) - \varphi(y)\|_{\mathcal{H}_k},$$

for $x, y \in \mathcal{X}$. We can perform the same construction to construct a metric on $\mathscr{P}(\mathcal{X})$ using Φ, which gives us the definition of the MMD (Gretton et al., 2012a):

$$MMD(k)(\mu, \nu) = \|\Phi(\mu) - \Phi(\nu)\|_{\mathcal{H}_k}.$$

The MMD can also be seen as arising from a lifting of kernels on $\mathcal{X}$ into kernels on $\mathscr{P}(\mathcal{X})$ (Guilbart, 1979). Given a kernel k on $\mathcal{X}$, define $K(\mu, \nu)$ for $\mu, \nu \in \mathscr{P}(\mathcal{X})$ as

$$K(\mu, \nu) = \langle \Phi(\mu), \Phi(\nu) \rangle_{\mathcal{H}_k} = \int_{\mathcal{X} \times \mathcal{X}} k(x, y)\, d(\mu \otimes \nu)(x, y).$$

It is immediate that K retains all properties of being a positive definite kernel as it arises from the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$. The MMD can then be seen as the metric ρ_K on $\mathscr{P}(\mathcal{X})$. We remark that the MMD with K allows one to metrize $\mathscr{P}(\mathscr{P}(\mathcal{X}))$, but we do not need this in this thesis.

[2] The Bochner integral is the extension of the Lebesgue integral to functions which take values in arbitrary Banach spaces, defined in the same way as the limit of simple functions.

One can also show that the MMD is an integral probability metric, as defined in subsection 2.1.3, since we can show that

$$MMD(k)(\mu,\nu) = \sup_{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\leq 1} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right| .$$

To see that this corresponds to the MMD as defined earlier, one can write out

$$\begin{aligned}\sup_{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\leq 1} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right| &= \sup_{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\leq 1} |\langle f, \Phi(\mu)\rangle_{\mathcal{H}_k} - \langle f, \Phi(\nu)\rangle_{\mathcal{H}_k}| \\ &= \sup_{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\leq 1} |\langle f, \Phi(\mu) - \Phi(\nu)\rangle_{\mathcal{H}_k}| \\ &= \|\Phi(\mu) - \Phi(\nu)\|_{\mathcal{H}_k},\end{aligned}$$

where we used the following fact for general Hilbert spaces $\mathcal{H}$: $\sup_{x:\|x\|_{\mathcal{H}}\leq 1}\langle x, y\rangle_{\mathcal{H}} = \|y\|_{\mathcal{H}}$, which follows from the Cauchy-Schwarz inequality.

4.1.3 The energy distance and semimetrics of negative type

A *semimetric* is a distance function which respects all metric axioms save for the triangle inequality. A semimetric space $(\mathcal{X}, \rho)$ is of *negative type* if for all $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$, we have

$$\sum_{i=1}^n \sum_{j=1}^n c_i c_j \rho(x_i, x_j) \leq 0.$$

Given a semimetric of negative type ρ on $\mathcal{X}$, we can define a distance on $\mathscr{P}(\mathcal{X})$ known as the *energy distance*, defined as

$$\mathcal{E}(\rho)(\mu,\nu) = \mathop{\mathbb{E}}_{x\sim\mu, y\sim\nu}[d(x,y)] - \frac{1}{2}\left(\mathop{\mathbb{E}}_{x_1,x_2\sim\mu}[d(x_1,x_2)] + \mathop{\mathbb{E}}_{y_1,y_2\sim\nu}[d(y_1,y_2)] \right).$$

The negative type of ρ is what guarantees that we have $\mathcal{E}(\rho)(\mu,\nu) \geq 0$ for all μ, ν. Metrics of negative type have a connection to positive definite kernels (Sejdinovic et al., 2013), in the sense that each kernel k induces a semimetric of negative type ρ_k through $\rho_k(x, y) = k(x, x) + k(y, y) - 2k(x, y)$. Conversely, a semimetric of negative type ρ induces a family of positive definite kernels K_ρ parametrised by a chosen base point $x_0 \in \mathcal{X}$:

$$K_\rho = \left\{ \frac{1}{2}(\rho(x, x_0) + \rho(x', x_0) - \rho(x, x')) : x_0 \in \mathcal{X} \right\}.$$

The relationship is symmetric, so that each kernel $k \in K_\rho$ has ρ as its induced semimetric. With this symmetry in mind, we call a kernel k and a semimetric of negative type an *equivalent pair* if they induce one another through the above construction. This equivalence does not only live in $\mathcal{X}$ however, as the following proposition shows that it lifts into $\mathscr{P}(\mathcal{X})$ as well.

Proposition 4.1.1. *Let (k, ρ) be an equivalent pair, and let $\mu, \nu \in \mathscr{P}(\mathcal{X})$. Then we have the equivalence*

$$MMD^2(k)(\mu, \nu) = \mathcal{E}(\rho)(\mu, \nu).$$

4.2 Behavioural kernels on Markov decision processes

Given an MDP $M = (\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, the state similarity metrics encountered can be described by the following form: for $x, y \in \mathcal{X}$

$$d(x, y) = d_1(x, y) + \gamma d_2(d)(\mathcal{P}(x), \mathcal{P}(y)),$$

where d_1 is a distance measure on $\mathcal{X}$, so that $d_1(x, y)$ represents the *immediate behavioural distance* of the states x and y, and d_2 lifts a distance on $\mathcal{X}$ into a distance on $\mathscr{P}(\mathcal{X})$, so that $d_2(d)(\mathcal{P}(x), \mathcal{P}(y))$ captures the *long-term behavioural distance* of the states.

In this chapter we take a similar approach, except rather than studying metrics, which measure the distance between two states, we will study positive definite kernels, which measure the similarity between two states (Aronszajn, 1950).

Extending this idea, we can define a behavioural similarity kernel on M to take a similar form. It should be the sum of a base kernel k_1 which measures the *immediate behavioural similarity* between two states, and a discounted mapping k_2 which *lifts* the kernel k into a kernel on $\mathscr{P}(\mathcal{X})$ which measures the *long-term behavioural similarity* of the states. Putting these together into a formular, we have

$$k(x, y) = k_1(x, y) + \gamma k_2(k)(\mathcal{P}(x), \mathcal{P}(y)).$$

A possible candidate for the immediate similarity is 1 minus the immediate distance in rewards, that is $1-|r_x^\pi - r_y^\pi|$, which is positive as we have 1-boundedness of the rewards. To measure the similarity of the transition kernels, we can use the method of lifting kernels into kernels onto probability distributions described by Guilbart (1979). Putting these together, we can look for a kernel of the form

$$k(x, y) = \big(1 - |r_x^\pi - r_y^\pi|\big) + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^\pi, y' \sim \mathcal{P}_y^\pi}[k(x', y')].$$

To see whether such a kernel actually exists, we can construct a sequence of iterates on the space of kernels.

We first discuss some preliminaries regarding the set of kernels on a space. Let $\mathscr{K}(\mathcal{X})$ be the set of positive definite kernels on $\mathcal{X}$. We have that $\mathscr{K}(\mathcal{X})$ is a subset of $\mathcal{B}(\mathcal{X} \times \mathcal{X})$, the set of real bounded functions on $\mathcal{X} \times \mathcal{X}$, which is complete under the $\|\cdot\|_\infty$ norm. Hence to show that $\mathscr{K}(\mathcal{X})$ is complete, it suffices to show that it is closed in $\mathcal{B}(\mathcal{X} \times \mathcal{X})$. We can now consider a sequence $\{k_i\}_{i\geq 1}$ in $\mathscr{K}(\mathcal{X})$ which converges to $k \in \mathcal{B}(\mathcal{X} \times \mathcal{X})$ in $\|\cdot\|_\infty$ and show that $k \in \mathscr{K}(\mathcal{X})$. This is equivalent to showing that k is both symmetric and positive definite, which follows immediately from the fact that each k_i is and the convergence is uniform. Hence $\mathscr{K}(\mathcal{X})$ is closed.

We can now define an operator $T_K^\pi : \mathscr{K}(\mathcal{X}) \to \mathscr{K}(\mathcal{X})$ as

$$T_K^\pi(k)(x, y) = \big(1 - |r_x^\pi - r_y^\pi|\big) + \gamma \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_x^\pi, y' \sim \mathcal{P}_y^\pi}[k(x', y')].$$

The fact that T_K^π indeed maps $\mathscr{K}(\mathcal{X})$ to $\mathscr{K}(\mathcal{X})$ follows from the previous paragraph describing that each operator is a kernel, and that the sum of two kernels is a kernel (Aronszajn, 1950). We can also see that T_K^π is a contraction with modulus γ in $\|\cdot\|_\infty$, as for $k, k' \in \mathscr{K}(\mathcal{X})$ we have that

$$\begin{aligned}
\|T_K^\pi(k) - T_K^\pi(k')\|_\infty &= \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} |T_K^\pi(k)(x,y) - T_K^\pi(k')(x,y)| \\
&= \gamma \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} \left| \int_{\mathcal{X}\times\mathcal{X}} k \, d(\mathcal{P}_x^\pi \otimes \mathcal{P}_y^\pi) - \int_{\mathcal{X}\times\mathcal{X}} k' \, d(\mathcal{P}_x^\pi \otimes \mathcal{P}_y^\pi) \right| \\
&= \gamma \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} \left| \int_{\mathcal{X}\times\mathcal{X}} (k - k') \, d(\mathcal{P}_x^\pi \otimes \mathcal{P}_y^\pi) \right| \\
&\leq \gamma \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} \left| \int_{\mathcal{X}\times\mathcal{X}} d(\mathcal{P}_x \otimes \mathcal{P}_y) \right| ||k - k'||_\infty \\
&= \gamma ||k - k'||_\infty .
\end{aligned}$$

Hence we have that T_K^π is a contraction. Combining this with the fact that $\mathscr{K}(\mathcal{X})$ is complete, we can use Banach's fixed point theorem to see the existence of a unique fixed point k^π satisfying $k^\pi = T_k^\pi(k^\pi)$.

Having a kernel on our MDP now gives us an RKHS of functions on the MDP, as well as an embedding of each state into said RKHS given by $\varphi(x) = k^\pi(x, \cdot)$, for $x \in \mathcal{X}$. As the kernel was built using a behavioural similarity recurrence, one can ask whether the Hilbert space distance between embeddings corresponds in any way to a behavioural distance between states. For states x and y, we can define a distance ρ_π as the Hilbert space distance between their embeddings:

$$\rho_\pi(x, y) = \|\varphi(x) - \varphi(y)\|_{\mathcal{H}_{k^\pi}} .$$

We refer to ρ_π as the *kernel similarity metric*. The squared Hilbert space norm can let us expand the above distance:

$$\begin{aligned}
\rho_\pi^2(x,y) &= \|\varphi(x)-\varphi(y)\|^2_{\mathcal{H}_{k^\pi}} \\
&= \langle \varphi(x)-\varphi(y), \varphi(x)-\varphi(y)\rangle_{\mathcal{H}_{k^\pi}} \\
&= \langle \varphi(x), \varphi(x)\rangle_{\mathcal{H}_{k^\pi}} + \langle \varphi(y), \varphi(y)\rangle_{\mathcal{H}_{k^\pi}} - 2\langle \varphi(x), \varphi(y)\rangle_{\mathcal{H}_{k^\pi}} \\
&= k^\pi(x,x) + k^\pi(y,y) - 2k^\pi(x,y) \\
&= |r_x^\pi - r_y^\pi| + \gamma\langle \Phi(\mathcal{P}_x^\pi, \mathcal{P}_x^\pi)\rangle_{\mathcal{H}_{k^\pi}} + \gamma\langle \Phi(\mathcal{P}_y^\pi, \mathcal{P}_y^\pi)\rangle_{\mathcal{H}_{k^\pi}} - 2\gamma\langle \Phi(\mathcal{P}_x^\pi, \mathcal{P}_y^\pi)\rangle_{\mathcal{H}_{k^\pi}} \\
&= |r_x^\pi - r_y^\pi| + \gamma MMD^2(k^\pi)(\mathcal{P}_x^\pi, \mathcal{P}_y^\pi). \qquad (4.0)
\end{aligned}$$

We can now see that the squared Hilbert space distance takes a familiar form to the behavioural metrics considered previously, using the squared MMD distance as the distance measure on the transition distributions.

4.2.1 Equivalence with reduced MICo distance

We now present a theorem that unifies the reduced MICo distance of section 3.3 with the kernel similarity metric as defined above. This is a significant result, as it demonstrates that ΠU^π exhibits a rich Hilbert space structure, while it had few known properties without this equivalence.

Theorem 4.2.1. *For any $x, y \in \mathcal{X}$, we have that*

$$\rho_\pi^2(x,y) = \Pi U^\pi(x,y).$$

Proof. To begin, we will make use of the sequences $(k_n)_{n\geq 0}$, $(U_n)_{n\geq 0}$ defined by $k_n \equiv 0$, $k_{n+1} = T_K^\pi(k_n)$, $U_n \equiv 0$, $U_{n+1} = T_M^\pi(U_n)$. Since both T_K^π and T_M^π are contractions, we know that $k_n \to k^\pi$ and $U_n \to U^\pi$ uniformly. To prove the statement, we will show that for all

$n \geq 0$ and $x, y \in \mathcal{X}$, we have that

$$k_n(x,x) + k_n(y,y) - 2k_n(x,y) = U_n(x,y) - \frac{1}{2}\left(U_n(x,x) + U_n(y,y)\right).$$

This combined with the fact that both sequences converge uniformly will allow us to take limits and finish the proof.

We will first begin with a necessary technical result, and show that for any measures μ, ν, and $n \geq 0$, we have that

$$\mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mu \\ y_1,y_2\sim\nu}}\left[k_n(x_1,x_2) + k_n(y_1,y_2) - 2k_n(x_1,y_1)\right] = \mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mu \\ y_1,y_2\sim\nu}}\left[U_n(x_1,y_1) - \frac{1}{2}(U_n(x_1,x_2) + U_n(y_1,y_2))\right].$$

We proceed to show this by induction. The base case is straightforward, as both sides are identically zero. We can now assume the induction hypothesis, and can write out

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mu \\ y_1,y_2\sim\nu}}\left[k_{n+1}(x_1,x_2) + k_{n+1}(y_1,y_2) - 2k_{n+1}(x_1,y_1)\right] \\
&= \mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mu \\ y_1,y_2\sim\nu}}\left[\left(|r^\pi_{x_1} - r^\pi_{y_1}| - \frac{1}{2}(|r^\pi_{x_1} - r^\pi_{x_2}| + |r^\pi_{y_1} - r^\pi_{y_2}|)\right) + \mathop{\mathbb{E}}_{\substack{x'_1\sim\mathcal{P}^\pi_{x_1} \\ x'_2\sim\mathcal{P}^\pi_{x_2} \\ y'_1\sim\mathcal{P}^\pi_{y_1} \\ y'_2\sim\mathcal{P}^\pi_{y_2}}}\left[k_n(x'_1,x'_2) + k_n(y'_1,y'_2) - 2k_n(x'_1,y'_1)\right]\right] \\
&= \mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mu \\ y_1,y_2\sim\nu}}\left[\left(|r^\pi_{x_1} - r^\pi_{y_1}| - \frac{1}{2}(|r^\pi_{x_1} - r^\pi_{x_2}| + |r^\pi_{y_1} - r^\pi_{y_2}|)\right) + \mathop{\mathbb{E}}_{\substack{x'_1,x'_2\sim\mathcal{P}^\pi_{x_2} \\ y'_1,y'_2\sim\mathcal{P}^\pi_{y_2}}}\left[k_n(x'_1,x'_2) + k_n(y'_1,y'_2) - 2k_n(x'_1,y'_1)\right]\right] \\
&= \mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mu \\ y_1,y_2\sim\nu}}\left[\left(|r^\pi_{x_1} - r^\pi_{y_1}| - \frac{1}{2}(|r^\pi_{x_1} - r^\pi_{x_2}| + |r^\pi_{y_1} - r^\pi_{y_2}|)\right) + \mathop{\mathbb{E}}_{\substack{x'_1,x'_2\sim\mathcal{P}^\pi_{x_2} \\ y'_1,y'_2\sim\mathcal{P}^\pi_{y_2}}}\left[U_n(x'_1,y'_1) - \frac{1}{2}(U_n(x'_1,x'_2) + U_n(y'_1,y'_2))\right]\right] \\
&= \mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mu \\ y_1,y_2\sim\nu}}\left[U_{n+1}(x_1,y_1) - \frac{1}{2}(U_{n+1}(x_1,x_2) + U_{n+1}(y_1,y_2)\right].
\end{aligned}$$

We can now conclude that for all n,

$$k_n(x,x) + k_n(y,y) - 2k_n(x,y) = U_n(x,y) - \frac{1}{2}(U_n(x,x) + U_n(y,y)),$$

as we can write out

$$\begin{aligned}
k_n(x,x) + k_n(y,y) - 2k_n(x,y) &= |r_x^\pi - r_y^\pi| + \gamma \mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mathcal{P}_x^\pi \\ y_1,y_2\sim\mathcal{P}_y^\pi}} [k_n(x_1,x_2) + k_n(y_1,y_2) - 2k_n(x_1,y_1)] \\
&= |r_x^\pi - r_y^\pi| + \gamma \mathop{\mathbb{E}}_{\substack{x_1,x_2\sim\mathcal{P}_x^\pi \\ y_1,y_2\sim\mathcal{P}_y^\pi}} \left[U_n(x_1,y_1) - \frac{1}{2}(U_n(x_1,x_2) + U_n(y_1,y_2))\right] \\
&= U_n(x,y) - \frac{1}{2}(U_n(x,x) + U_n(y,y)).
\end{aligned}$$

Since $(k_n)_{n\geq 0}$ and $(U_n)_{n\geq 0}$ both converge uniformly, we can take limits and conclude that

$$\rho_\pi^2(x,y) = k^\pi(x,x) + k^\pi(y,y) - 2k^\pi(x,y) = U^\pi(x,y) - \frac{1}{2}(U^\pi(x,x) + U^\pi(y,y)) = \Pi U^\pi(x,y).$$

□

4.3 Alternative parametrisations of the reduced MICo

We recall from subsection 3.4.2 that the parametrisation of the reduced MICo in the neural network took a peculiar form, as we used

$$U_\omega(x,y) = \frac{\|\phi_\omega(x)\| + \|\phi_\omega(y)\|}{2} + \beta\,\theta(\phi_\omega(x), \phi_\omega(y)),$$

so that $\Pi U_\omega \approx \beta\theta(\phi_\omega(x), \phi_\omega(y))$. Although this parametrisation achieves its goal of representing the reduced MICo distance, it lacks interpretability and its motivation may be opaque. Using the equivalence from Theorem 4.2.1, we can propose more transparent and intuitive parametrisations.

Following the notation in subsection 3.4.2, we wish to approximate the kernel k^π using a parametrised k_ω. To represent k_ω in the neural network, we can use one of the commonly used kernels on $\mathbb{R}^d$ (Shawe-Taylor and Cristianini, 2004), such as:

- Linear kernel: $k_\omega(x, y) = \langle \phi_\omega(x), \phi_\omega(y) \rangle$
- Polynomial kernel: $k_\omega(x, y) = (\langle \phi_\omega(x), \phi_\omega(y) \rangle + C)^n$, where C and n are hyperparameters
- Gaussian kernel: $k_\omega(x, y) = e^{-\frac{\|\phi_\omega(x) - \phi_\omega(y)\|^2}{2\sigma^2}}$, where σ is a hyperparameter,

to name a few. Once a parametrisation k_ω is chosen, one can learn $k_\omega \approx k^\pi$ using the following loss:

$$\mathcal{L}_{\text{Kernel}}(\omega) = \mathop{\mathbb{E}}_{\langle x, r_x, x' \rangle, \langle y, r_y, y' \rangle \sim \mathcal{D}} \left[\left(k_\omega(x, y) - \big(1 - |r_x - r_y| + \gamma k_{\bar{\omega}}(x', y')\big) \right)^2 \right],$$

which can then be incorporated into the loss of any agent in the same process done in subsection 3.4.2. We then have that the equivalent distance to k_ω, d_ω, satisfies

$$\Pi U^\pi(x, y) = \rho_\pi^2(x, y) \approx d_\omega(x, y).$$

This is a much more flexible parametrisation than the one presented in subsection 3.4.2, importantly any distance can be used rather than the angular distance (to choose a given distance, one must simply choose its equivalent kernel). We hypothesize that different kernels will be better across different MDPs and environments.

4.4 Discussion and future work

In this chapter, we introduce a novel perspective on state similarity metric learning; rather than directly learning distance functions, we learned similarity functions instead. This allowed us to leverage theory from reproducing kernel Hilbert spaces, and prove equivalence to the reduced MICo distance we introduced in Chapter 3.

The first direction for future work is empirical, as section 4.3 provides many interesting directions to explore. This direction is interesting for a number of reasons: firstly, subsection 3.4.3 demonstrates very strong performance, and it is interesting to see if one of the new parametrisations can achieve even greater state of the art performance. Secondly, it can be interesting to find which kernel parametrisation is best for various MDPs, depending on the properties. In particular, different kernels can exploit various symmetries or invariances in the underlying environment. The choice of what kernel to use is known as the *kernel choice problem*, and has been studied extensively in machine learning (Muandet et al., 2017; Sriperumbudur et al., 2009a; Gretton et al., 2012b).

A second direction for future work is one of a more theoretical nature. When learning a state metric d to approximate the distance between embeddings $\|\phi(x) - \phi(y)\|$, we are approximating d using a Hilbert space distance (the Euclidean distance). As this chapter shows, ΠU^π is a squared Hilbert space distance. On the other hand, the Kantorovich is not a Hilbertian metric (Gehér et al., 2021), and so in particular the bisimulation metric $d^\pi_\sim$ is not Hilbertian, meaning there does not exist a Hilbert space with embedding φ such that

$$d^\pi_\sim(x, y) = \|\varphi(x) - \varphi(y)\|_{\mathcal{H}}.$$

The direction for research is to investigate what effects this phenomenon has on the learnability of the metrics, does it imply that naturally ΠU^π can be approximated in Euclidean space better (i.e. with less distortion) than $d^\pi_\sim$? In this vein, a number of previous works have learnt $d^\pi_\sim$ in neural networks (Castro, 2020; Zhang et al., 2021; Kemertas and Aumentado-Armstrong, 2021; Kemertas and Jepson, 2022), however none have studied the quality of the learnt metric, through distortion or other criteria.

Chapter 5

A Distributional Perspective

In the behavioural distances seen thus far, the immediate distance between two states x and y is their distance in expected rewards, that is $|\mathbb{E}[\mathcal{R}_x] - \mathbb{E}[\mathcal{R}_y]|$ (we are intentionally vague about whether we are considering a single action, $\mathcal{R}^a$, or for a given policy $\mathcal{R}^\pi$, since this discussion holds for both settings). In the convergence theorems we have seen, we require an assumption on the rewards from a state being deterministic, such as in Theorem 3.2.6. We will now see that this is due to a deeper issue regarding sampling the absolute difference of expectations of random variables.

Let X and Y be two real-valued random variables on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We are interested in estimating the quantity $|\mathbb{E}[X] - \mathbb{E}[Y]|$, given samples $\{x_1, \ldots, x_n\}$ from X and $\{y_1, \ldots, y_n\}$ from Y. A common estimator for this is

$$\frac{1}{n^2}\left|\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - y_j)\right|.$$

This estimator is consistent, and of near-minimal variance, however it is biased in general:

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - y_j)\right| &\geq \left|\mathbb{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - y_j)\right)\right| \\ &= \frac{1}{n^2}\left|\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbb{E}(X) - \mathbb{E}(Y)\right)\right| \\ &= \left|\mathbb{E}(X) - \mathbb{E}(Y)\right|,\end{aligned}$$

where we used Jensen's inequality. This is not sufficient to show bias however, since the inequality is not strict. To do this, we can use the fact that Jensen's inequality attains equality under two possible conditions: the outer function is affine, or the inner random variable is almost surely constant. In this case, the outer function $z \mapsto |z|$ is not affine, so we have equality if and only if X and Y are both almost surely constant. This is now sufficient to answer our question regarding the bias: this estimator is unbiased if X and Y are both almost surely constant, and biased otherwise. Although this analysis was done for a particular estimator, this phenomenon is true in general: for general random variables X and Y, there is no unbiased estimator for $\mathbb{E}[|X - Y|]$ (Elandt, 1961).

The above paragraph provides us another view of why we required the reward to only depend on the state for Theorem 3.2.6 - in this case, $\mathcal{R}_x$ and $\mathcal{R}_y$ are almost surely constant random variables, and thus $|\,\mathbb{E}[\mathcal{R}_x] - \mathbb{E}[\mathcal{R}_y]|$ can be estimated without bias. We recall that for stochastic approximation, having an unbiased estimator of our target is imperative (Bertsekas and Tsitsiklis, 1996). With this in mind, it seems that the assumptions made in Theorem 3.2.6 are essentially the best we can hope to do, since Elandt (1961) shows that it is not possible for general rewards.

5.1 A formulation for general rewards

Instead of considering assumptions under which we can learn $|\mathbb{E}[\mathcal{R}_x] - \mathbb{E}[\mathcal{R}_y]|$, we can ask what we instead learn in the general case. That is, if we perform the online updates as in Theorem 3.2.6, do we converge to a distance, and if so to which?

Given samples x, y from random variables X and Y, the value $|x - y|$ is an unbiased estimate of $\mathbb{E}[|X - Y|]$. With this in mind, we will now consider an adaptation of the theory from Chapters 2 and 3, but using $\mathbb{E}[|\mathcal{R}_x - \mathcal{R}_y|]$ instead of $|\mathbb{E}[\mathcal{R}_x] - \mathbb{E}[\mathcal{R}_y]|$.

We begin with an adaptation of the MICo operator. Let us define $T^\pi_{\bar{U}} : \mathcal{M}^{diff} \to \mathcal{M}^{diff}$ as

$$T^\pi_{\bar{U}}(U)(x, y) = \mathbb{E}\left[|\mathcal{R}^\pi_x - \mathcal{R}^\pi_y|\right] + \gamma\, d_{LK}(U)(\mathcal{P}^\pi_x, \mathcal{P}^\pi_y).$$

It is a straightforward adaptation of the proof of Theorem 3.2.4 to see that $T^\pi_{\bar{U}}$ has a unique fixed point, $\bar{U}^\pi$.

From Jensen's inequality, we know that $|\mathbb{E}[X] - \mathbb{E}[Y]| \leq \mathbb{E}[|X - Y|]$, and so in particular we have that $U^\pi(x, y) \leq \bar{U}^\pi(x, y)$ for all $x, y \in \mathcal{X}$. From this we can see that for any $x, y \in \mathcal{X}$ we have $|V^\pi(x) - V^\pi(y)| \leq \bar{U}^\pi(x, y)$, since

$$|V^\pi(x) - V^\pi(y)| \leq U^\pi(x, y) \leq \bar{U}^\pi(x, y).$$

However, we can say more than this. Let $\eta^\pi(x)$ be the distribution of $\sum_{t \geq 0} \gamma^t \mathcal{R}_t$ under π, given that $X_0 = x$. Then $\eta^\pi(x)$ is a measure on $\mathbb{R}$, with expected value $V^\pi(x)$. $\eta^\pi(x)$ is known as the *return distribution*, and is a central concept in distributional reinforcement learning (Bellemare et al., 2017, 2022).

Using Jensen's inequality, we can see that

$$|V^\pi(x) - V^\pi(y)| \leq \mathop{\mathbb{E}}_{\substack{G^\pi(x) \sim \eta^\pi(x) \\ G^\pi(y) \sim \eta^\pi(y)}} \left[|G^\pi(x) - G^\pi(y)|\right].$$

We now show that the quantity on the right hand side is also bounded by $\bar{U}^\pi$.

Theorem 5.1.1. *For any $x, y \in \mathcal{X}$, we have*

$$\mathop{\mathbb{E}}_{\substack{G^\pi(x)\sim\eta^\pi(x)\\ G^\pi(y)\sim\eta^\pi(y)}} [|G^\pi(x) - G^\pi(y)|] \leq \bar{U}^\pi(x, y).$$

Proof. To prove this, we will first review some necessary concepts in distributional reinforcement learning. We will let $\mathscr{P}^{\mathcal{X}}$ be the set of return distribution functions which assign a measure over $\mathbb{R}$ for each $x \in \mathcal{X}$. The supremum Kantorovich distance is used to metrize $\mathscr{P}^{\mathcal{X}}$, and is defined as

$$\bar{\mathcal{W}}_1(\eta, \eta') = \sup_{x\in\mathcal{X}} \mathcal{W}_1(\eta(x), \eta'(x)),$$

for $\eta, \eta' \in \mathscr{P}^{\mathcal{X}}$. Using this, the metric space $(\mathscr{P}^{\mathcal{X}}, \bar{\mathcal{W}}_1)$ is complete. The *distributional Bellman operator* (Rowland et al., 2018; Bellemare et al., 2022) is the map $\mathcal{T}^\pi : \mathscr{P}^{\mathcal{X}} \to \mathscr{P}^{\mathcal{X}}$ given by

$$(\mathcal{T}^\pi\eta)(x) = \mathbb{E}_\pi[(b_{R,\gamma})_\#\eta(X_1) \,\big|\, X_0 = x],$$

where $b_{r,\gamma}(x) = r + \gamma x$. $\mathcal{T}^\pi$ is a contraction with modulus γ in $(\mathscr{P}^{\mathcal{X}}, \bar{\mathcal{W}}_1)$, hence one can apply Banach's fixed point theorem and refer to the unique fixed point as η^π. Moreover for any $\eta_0 \in \mathscr{P}^{\mathcal{X}}$, one can define a sequence as $\eta_{k+1} = \mathcal{T}^\pi\eta_k$, which will converge to η^π.

With this in mind, let $\{\eta_k\}_{k\geq 0}$ and $\{\bar{U}_k\}_{\geq 0}$ be sequences defined by $\eta_0(x) = \delta_0$ for all $x \in \mathcal{X}$, $\bar{U}_0(x, y) = 0$ for all $x, y \in \mathcal{X}$, $\eta_{k+1} = \mathcal{T}^\pi\eta_k$, and $\bar{U}_{k+1} = T^\pi_{\bar{U}}\bar{U}_k$. Then we know that as $k \to \infty$, we have that both $\eta_k \to \eta^\pi$ and $\bar{U}_k \to \bar{U}^\pi$.

We will now show by induction that for all k, we have that for any $x, y \in \mathcal{X}$,

$$\mathop{\mathbb{E}}_{G^k(x)\sim\eta_k(x), G^k(y)\sim\eta_k(y)} \left[|G^k(x) - G^k(y)|\right] \leq \bar{U}_k(x, y).$$

Let us fix $x, y \in \mathcal{X}$ arbitrarily. In the base case, we have that $\eta_0(x) = \eta_0(y) = \delta_0$, and $G^k(x)$ and $G^k(y)$ are both 0 almost surely, so the left hand side is 0. Similarly, the right hand side is 0 by definition of $\bar{U}_0$.

Let us write $(x, A, \mathcal{R}_x, X')$ and $(y, A, \mathcal{R}_y, Y')$ to be independent sample transitions from x and y. We use the fact that for all k, $\mathcal{R}_x + \gamma G^k(X')$ has distribution $\eta_{k+1}(x)$, and $\mathcal{R}_y + \gamma G^k(Y')$ has distribution $\eta_{k+1}(y)$ (Bellemare et al., 2022). We can write out

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{G^{k+1}(x)\sim\eta_{k+1}(x)\\ G^{k+1}(y)\sim\eta_{k+1}(y)}}\left[\left|G^k(x) - G^k(y)\right|\right] &= \mathop{\mathbb{E}}_{\substack{\mathcal{R}_x+\gamma G^k(X')\sim\mathcal{T}^\pi\eta_k(x)\\ \mathcal{R}_y+\gamma G^k(Y')\sim\mathcal{T}^\pi\eta_k(y)}}\left[\left|(\mathcal{R}_x + \gamma G^k(X')) - (\mathcal{R}_y + \gamma G^k(Y'))\right|\right] \\
&\leq \mathop{\mathbb{E}}_{\substack{\mathcal{R}_x+\gamma G^k(X')\sim\mathcal{T}^\pi\eta_k(x)\\ \mathcal{R}_y+\gamma G^k(Y')\sim\mathcal{T}^\pi\eta_k(y)}}\left[\left|\mathcal{R}_x - \mathcal{R}_y\right| + \gamma\left|G^k(X') - G^k(Y')\right|\right] \\
&= \mathbb{E}_\pi\left[\left|\mathcal{R}_x - \mathcal{R}_y\right|\right] + \gamma \mathop{\mathbb{E}}_{x'\sim\mathcal{P}^\pi_x, y'\sim\mathcal{P}^\pi_y}\left[\mathop{\mathbb{E}}_{\substack{G^k(X')\sim\eta_k(x')\\ G^k(Y')\sim\eta_k(y')}}\left[\left|G^k(X') - G^k(Y')\right|\right]\right] \\
&\leq \mathbb{E}_\pi\left[\left|\mathcal{R}_x - \mathcal{R}_y\right|\right] + \mathop{\mathbb{E}}_{x'\sim\mathcal{P}^\pi_x, y'\sim\mathcal{P}^\pi_y}\left[\bar{U}_k(x', y')\right] \\
&= \bar{U}_{k+1}(x, y),
\end{aligned}$$

where we used the induction hypothesis in the second to last line.

Since convergence is uniform, we can take $k \to \infty$ to see that

$$\mathop{\mathbb{E}}_{G^\pi(x)\sim\eta^\pi(x), G^\pi(y)\sim\eta^\pi(y)}\left[\left|G^\pi(x) - G^\pi(y)\right|\right] \leq \bar{U}^\pi(x, y).$$

□

We now show that this is a novel property of $\bar{U}^\pi$, and it does not hold for the original MICo distance U^π.

Proposition 5.1.2. *In general, we do not have that*

$$\mathop{\mathbb{E}}_{G^\pi(x)\sim\eta^\pi(x), G^\pi(y)\sim\eta^\pi(y)}\left[\left|G^\pi(x) - G^\pi(y)\right|\right] \leq U^\pi(x, y)$$

for all states x and y.

Proof. Let us consider an MDP with two states and a single action, $\mathcal{X} = \{x, y\}$, $\mathcal{A} = \{a\}$, $\mathcal{P}^a_x(x) = 1$, $\mathcal{P}^a_y(y) = 1$, $\mathcal{P}_\mathcal{R}(\cdot \mid x, a) = \delta_0$,$\mathcal{P}_\mathcal{R}(\cdot \mid y, a) = \frac{1}{2}(\delta_{-1} + \delta_1)$, and to simplify the analysis let us take $\gamma = 0$. As there is a single action, there exists a single policy π. As the discount factor is 0, the return distribution at a state is simply the immediate reward distribution, so we have that $\eta^\pi(x) = \delta_0$, $\eta^\pi(y) = \frac{1}{2}(\delta_{-1} + \delta_1)$. It is straightforward to see that $V^\pi(x) = V^\pi(y) = 0$. Moreover, since we have $r^\pi_x = r^\pi_y = 0$ and both states transition to themselves with probability 1, we see that $U^\pi(x, y) = 0$. However, we can calculate that

$$\mathop{\mathbb{E}}_{G^\pi(x)\sim\eta^\pi(x), G^\pi(y)\sim\eta^\pi(y)}\left[|G^\pi(x) - G^\pi(y)|\right] = \frac{1}{2}(|1| + |-1|) = 1 > 0 = U^\pi(x, y).$$

□

We now show that what we claimed at the beginning of this section is true, that is that when following the update scheme from Theorem 3.2.6, we converge to $\bar{U}^\pi$ in general settings. This indicates that this is not necessarily a *new* distance, but instead the one which naturally arises during the learning procedure.

Theorem 5.1.3. *Let $M = (\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ be a Markov decision process, and π be any policy on M. Let $(X_k, A_k, R_k, X'_k)_{k\geq 0}$ and $(Y_k, A'_k, R'_k, Y'_k)_{k\geq 0}$ be two sequences of transitions in M following π. Moreover, let $(\alpha_k)_{k\geq 0}$ be a sequence of stepsizes satisfying the Robbins-Monro conditions. Let $(U_k)_{k\geq 0}$ be a sequence of estimates constructed by choosing U_0 arbitrarily, and setting*

$$\begin{aligned} U_{k+1}(X_k, Y_k) &= (1 - \alpha_k)\, U_k(X_k, Y_k) + \alpha_k\, \left(|R_k - R'_k| + \gamma U_k(X'_k, Y'_k)\right) \\ U_{k+1}(x, y) &= U_k(x, y) \text{ if } (x, y) \neq (X_k, Y_k). \end{aligned}$$

Then if each pair of states is sampled infinitely often, we have that $U_k \to U^\pi$ with probability 1.

Proof. The result mainly follows from the proof of Theorem 3.2.6, it only remains to show that $|R_k - R'_k|$ is an unbiased estimate of $\mathbb{E}[|\mathcal{R}^\pi_{X_K} - \mathcal{R}^\pi_{Y_k}|]$ for any rewards. But this is straightforward to see as

$$\begin{aligned} \mathbb{E}[|R_k - R'_k|] &= \mathbb{E}[|R_k - R'_k| \,\big|\, X_k, Y_k] \\ &= \mathbb{E}[|\mathcal{R}^\pi_{X_k} - \mathcal{R}^\pi_{Y_k}|], \end{aligned}$$

as we have $A_k \sim \pi(\cdot|X_k)$, $A'_k \sim \pi(\cdot|Y_k)$, so that $R_k \sim \mathcal{R}^\pi_{X_k}$, $R'_k \sim \mathcal{R}^\pi_{Y_k}$. □

5.2 Discussion and future work

In this chapter, we introduce a theory in which we replace the difference in expected rewards with the expected absolute distance in rewards. We then proved that this provided us an upper bound between the expected absolute difference of return distributions, which was not satisfied by U^π. Most importantly, we prove that when using the update scheme used to learn U^π, we converge to $\bar{U}^\pi$ in general settings. This points to $\bar{U}^\pi$ being the more natural distance than U^π, it is what is being learnt when we do not have assumptions on the environment. We now highlight two interesting directions for future work.

Firstly, there is an opportunity to critically reexamine the current literature. A number of papers (Castro et al., 2021; Zhang et al., 2021; Kemertas and Aumentado-Armstrong, 2021; Kemertas and Jepson, 2022) have considered learning bisimulation metrics in environments where the reward is stochastic, but their theories consider the difference in expected rewards, that is $|\mathbb{E}[\mathcal{R}^\pi_x] - \mathbb{E}[\mathcal{R}^\pi_y]|$. However, the proofs in this chapter can be applied to show that in all of these cases, the distance being learned is instead $\mathbb{E}[|\mathcal{R}^\pi_x - \mathcal{R}^\pi_y|]$. It is of interest to see how this affects the theory presented in these papers, in order to reduce the theory-practice gap. Related to this, it would be interesting to see how the representations differ when considering $\mathbb{E}[|\mathcal{R}^\pi_x - \mathcal{R}^\pi_y|]$ as compared to $|\mathbb{E}[\mathcal{R}^\pi_x] - \mathbb{E}[\mathcal{R}^\pi_y]|$. This can be done theoretically (i.e. by studying how generalization properties of one compares to the other), or done empirically, which would need to be done in small-scale environments where both can be learnt exactly.

Secondly, Theorem 5.1.1 is the first theorem in literature (at least known to the author) which provides a *risk-sensitive* connection to state-similarity metrics (where risk-sensitive indicates that we are considering more than just the means of the return distributions η^π). Distributional approaches to machine learning have been of great interest in recent years

(Bellemare et al., 2017; Dabney et al., 2018b; Bellemare et al., 2022), and have provided state of the art results in empirical settings (Dabney et al., 2018a; Hessel et al., 2018). It would be very interesting to see if $\bar{U}^{\pi}$ has any relationships to quantities of interest in distributional RL literature (such as the Kantorovich or Cramér distance between return distributions), or if concepts from distributional literature can be leveraged to produce new distances.

Chapter 6

A unifying framework

In the previous chapters, we have seen a number of different behavioural metrics on state spaces, with varying degrees of relation to one another. With so many variations, it may be difficult to choose which distance to use, as well as intuitively understand how two differ.

In this chapter, we present a framework through which all the previous metrics considered can be seen as arising from. This framework can be seen as an extension of the work of Ferns and Precup (2014), where they proved that the classical bisimulation metric in fact corresponds to the optimal value function of a particular MDP. We expand upon this, and show that all of the considered metrics can be constructed in the same fashion.

Given an MDP $M = (\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, there are a number of ways one can define an *auxiliary MDP* $\tilde{M} = (\tilde{\mathcal{X}}, \tilde{\mathcal{A}}, \tilde{\mathcal{P}}, \tilde{\mathcal{R}}, \gamma)$. We will always take $\tilde{\mathcal{X}} = \mathcal{X} \times \mathcal{X}$, and will always have $\tilde{\mathcal{P}}$ be a coupling of $\mathcal{P}$ with itself (whether we take couplings of transitions with the same action or with different actions depends on whether $\tilde{\mathcal{A}}$ is equal to $\mathcal{A}$ or $\mathcal{A} \times \mathcal{A}$). Given a coupling λ, we will write the policy (optimal) value function for the auxiliary MDP as $\tilde{V}^{\pi}_{\lambda}$ ($\tilde{V}^{*}_{\lambda}$), when the other parameters of the auxiliary MDP are known.

We then discuss convergence of learning various metrics under changing policies, and present a theorem on when this convergence occurs. Using the auxiliary MDP framework presented, we then prove that this convergence indeed occurs for all metrics considered.

6.1 State similarity metrics as value functions

We now begin an analysis of different auxiliary MDPs, and for each construction, we follow a similar recipe. First, we will choose the action space and the reward function, and we choose a set of couplings within with the transition distribution will live. We then consider two couplings, the independent coupling and the optimal coupling, which induce two different MDPs. We then consider the value functions of these MDPs (either optimal or policy), which in turn induce a distance on the original MDP.

6.1.1 Maximal action metrics

We will first take $\tilde{\mathcal{A}} = \mathcal{A}$ and look at transition functions of the form $\tilde{\mathcal{P}}^a_{(x,y)} \in \Lambda(\mathcal{P}^a_x, \mathcal{P}^a_y)$. We take the reward function to be $\tilde{\mathcal{R}}^a(x,y) = |r^a_x - r^a_y|$. We now consider the two couplings:

1. Independent coupling: One can take $\tilde{\mathcal{P}}^a_{(x,y)}$ to be the product distribution $\mathcal{P}^a_x \times \mathcal{P}^a_y$. The optimal value function $\tilde{V}^*$ can then be derived as

$$\begin{aligned}\tilde{V}^*((x,y)) &= \max_{a\in\mathcal{A}} \left(\tilde{\mathcal{R}}^a(x,y) + \gamma \mathop{\mathbb{E}}_{(x',y')\sim\tilde{\mathcal{P}}^a_{(x,y)}} \left[\tilde{V}^*((x',y')) \right] \right) \\ &= \max_{a\in\mathcal{A}} \left(|r^a_x - r^a_y| + \gamma \mathop{\mathbb{E}}_{x'\sim\mathcal{P}^a_x, y'\sim\mathcal{P}^a_y} \left[\tilde{V}^*((x',y')) \right] \right).\end{aligned}$$

   We can identify this as a distance U^*, written as

$$U^*(x,y) = \max_{a\in\mathcal{A}} \left(|r^a_x - r^a_y| + \gamma\, d_{LK}(\mathcal{P}^a_x, \mathcal{P}^a_y) \right).$$

   This can be seen as a ‘MICo’-like approach to the classical bisimulation, we take the maximum across all actions, but considering the Łukaszyk–Karmowski distance rather than the Kantorovich metric.

2. Optimal coupling: One can look at the coupling λ which minimizes $\tilde{V}^*$, formally that is $\lambda = \arg\min_{\lambda\in\Lambda(\mathcal{P}^a_x,\mathcal{P}^a_y)} \tilde{V}^*_\lambda$. This was the setting considered in Ferns and Precup (2014). In this case, the auxiliary optimal value function $\tilde{V}^*_\lambda$ corresponds to the

bisimulation metric $d^{\sim}$ in the original MDP, as

$$\tilde{V}^*((x,y)) = d^{\sim}(x,y) = \max_{a\in\mathcal{A}}\left(|r_x^a - r_y^a| + \gamma\mathcal{W}(d^*)(\mathcal{P}_x^a, \mathcal{P}_y^a)\right).$$

These two distances can be related to value functions in the original MDP through the following proposition:

Proposition 6.1.1. *For any $x, y \in \mathcal{X}$, we have that*

$$|V^*(x) - V^*(y)| \leq d^{\sim}(x,y) \leq U^*(x,y).$$

Proof. We know from (Ferns et al., 2004) that for any $x, y \in \mathcal{X}$ we have that

$$|V^*(x) - V^*(y)| \leq d^{\sim}(x,y),$$

and so it remains to show that we also have $d^{\sim}(x,y) \leq U^*(x,y)$. This can be seen through induction using the fact that for any random variables X and Y, one has

$$|\mathbb{E}[X] - \mathbb{E}[Y]| \leq \mathbb{E}[|X - Y|].$$

□

6.1.2 Action-independent rewards

We now consider the case where $\tilde{\mathcal{A}} = \mathcal{A}\times\mathcal{A}$, and we *fix* a policy π in the original MDP, and define the auxiliary reward as $\tilde{\mathcal{R}}^{(a,b)}(x,y) = |r_x^\pi - r_y^\pi|$. The subsection name comes from the fact that $\tilde{\mathcal{R}}$ has no dependence on the action (a,b) which was taken, and instead is fixed from π. Since we are looking at pairs of actions, the transition distributions we will be looking at are of the form $\tilde{\mathcal{P}}_{(x,y)}^{(a,b)} \in \Lambda(\mathcal{P}_x^a, \mathcal{P}_y^b)$. We will also be interested in the policy $\tilde{\pi}$ on $\tilde{M}$, which is defined as independently following π in each coordinate: $\tilde{\pi}(\cdot,\cdot|x,y) = \pi(\cdot|x)\times\pi(\cdot|y)$. We can now look at which metrics are produced in this setting.

- Independent coupling: One can take $\tilde{\mathcal{P}}^{(a,b)}_{(x,y)}$ to be the product distribution $\mathcal{P}^a_x \times \mathcal{P}^b_y$. The policy value function $\tilde{V}^{\tilde{\pi}}$ then corresponds to a distance U^π on the original MDP defined as

$$\tilde{V}^{\tilde{\pi}}((x,y)) = U^\pi(x,y) = |r^\pi_x - r^\pi_y| + \gamma d_{LK}(U^\pi)(\mathcal{P}^\pi_x, \mathcal{P}^\pi_y).$$

- Optimal coupling: Similarly to the previous section, one can consider the optimal coupling $\lambda = \arg\min_{\lambda\in\Lambda(\mathcal{P}^a_x,\mathcal{P}^b_y)} \tilde{V}^{\tilde{\pi}}_\lambda$. In this case, the policy value function $\tilde{V}^{\tilde{\pi}}_\lambda$ corresponds to the distance $d^\pi_\sim$ in the original MDP, defined as

$$\tilde{V}^{\tilde{\pi}}((x,y)) = d^\pi_\sim(x,y) = |r^\pi_x - r^\pi_y| + \gamma\mathcal{W}(d^\pi)(\mathcal{P}^\pi_x, \mathcal{P}^\pi_y).$$

One can now see that U^π corresponds to the original MICo distance, and d^π corresponds to on-policy bisimulation. As we know, these correspond to the policy-value functions since for any $x, y \in \mathcal{X}$,

$$|V^\pi(x) - V^\pi(y)| \leq d^\pi(x,y) \leq U^\pi(x,y).$$

6.1.3 Action-dependent rewards

In the previous section, we had the transition functions depend on actions, but the rewards did not. One can consider modifying the previous case to the setting where both depend on actions, so we take $\tilde{\mathcal{P}}^{(a,b)}_{(x,y)} \in \Lambda(\mathcal{P}^a_x, \mathcal{P}^b_y)$ and $\tilde{\mathcal{R}}^{(a,b)}(x,y) = |r^a_x - r^b_y|$. Then for any policy π on M, one can once again consider the policy $\tilde{\pi}$ on $\tilde{M}$ which follows π independently in each coordinate: $\tilde{\pi}(\cdot,\cdot|x,y) = \pi(\cdot|x) \times \pi(\cdot|y)$. An important difference between the previous subsection is this construction can support any policy π in the original MDP, while the previous construction only supported the fixed policy which was used in the construction. We can now see which metrics are induced.

- Independent coupling: One can take $\tilde{\mathcal{P}}^{(a,b)}_{(x,y)}$ to be the product distribution $\mathcal{P}^a_x \times \mathcal{P}^b_y$. The policy value function $\tilde{V}^{\tilde{\pi}}$ then corresponds to a distance $\bar{U}^\pi$ on the original MDP defined as

$$\tilde{V}^{\tilde{\pi}}((x,y)) = \bar{U}^{\pi}(x,y) = \mathop{\mathbb{E}}_{r_x\sim R^{\pi}_x, r_y\sim R^{\pi}_y}\left[|r_x - r_y|\right] + \gamma d_{LK}(\bar{U}^{\pi})(\mathcal{P}^{\pi}_x, \mathcal{P}^{\pi}_y).$$

- Optimal coupling: One can consider the optimal coupling $\lambda = \arg\min_{\lambda\in\Lambda(\mathcal{P}^a_x,\mathcal{P}^b_y)} \tilde{V}^{\tilde{\pi}}_{\lambda}$. In this case, the policy value function $\tilde{V}^{\tilde{\pi}}_{\lambda}$ corresponds to the distance $\bar{d}^{\pi}_{\sim}$ in the original MDP, defined as

$$\tilde{V}^{\tilde{\pi}}((x,y)) = \bar{d}^{\pi}_{\sim}(x,y) = \mathop{\mathbb{E}}_{r_x\sim R^{\pi}_x, r_y\sim R^{\pi}_y}\left[|r_x - r_y|\right] + \gamma \mathcal{W}(\bar{d}^{\pi})(\mathcal{P}^{\pi}_x, \mathcal{P}^{\pi}_y).$$

In this case, $\bar{U}^{\pi}$ corresponds to the modified MICo distance introduced in Chapter 5, and $\bar{d}^{\pi}$ follows from a similar modification of the on-policy bisimulation metric $d^{\pi}_{\sim}$. The connection to the original policy value functions can be seen as a straightforward modification of Theorem 5.1.1, for any $x, y \in \mathcal{X}$, we have

$$\mathop{\mathbb{E}}_{G^{\pi}_x\sim\eta^{\pi}(x), G^{\pi}_y\sim\eta^{\pi}(y)}\left[|G^{\pi}_x - G^{\pi}_y|\right] \leq \bar{d}^{\pi}(x,y) \leq \bar{U}^{\pi}(x,y).$$

6.2 Convergence under changing policies

Through the previous chapters and previous literature, the convergence analysis of state similarity metrics has followed a common recipe: fix a policy π, produce an operator for this policy which is a contraction, and conclude that iterating this operator produces a fixed point, which is our state similarity metric. This recipe however, is inconsistent with what is done in practice, where one has a changing policy as the learning progresses. That is, one may begin learning the metric with a crude, exploratory policy at the beginning of training, and continue learning the metric using a more refined, greedy policy near the end of training. The previous convergence results are not sufficient to address this setting.

We will now present a general convergence result regarding learning metrics when the underlying policy is changing, and prove that a number of metrics considered thus far satisfy the conditions required to satisfy this convergence result. The argument behind the proof of this theorem is an adaptation of the proofs found in section 6.2. of Bertsekas and Tsitsiklis (1996).

Definition 6.2.1. A sequence of policies $(\pi_k)_{k\geq 0}$ converges uniformly to π^* if we have that

$$\sup_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} |\pi_k(a\,|\,x) - \pi^*(a\,|\,x)| \to 0$$

as $k \to \infty$.

Theorem 6.2.2. *Suppose that $\mathcal{X}$ is finite, $|\mathcal{R}|$ is almost surely bounded by R_{max}, $(\pi_k)_{k\geq 0}$ is a sequence of policies converging uniformly to π^*, and let $(V, \|\cdot\|_\infty)$ be a space of functions with the supremum norm. Moreover, for each policy π let $F_d^\pi : V \to V$ be a contraction in $(V, \|\cdot\|_\infty)$ with fixed point d^π. Suppose that F_d is Lipschitz continuous with constant C in π at π^*, that is we have*

$$\|F_d^\pi - F_d^{\pi^*}\|_\infty \leq C \sup_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} |\pi(a\,|\,x) - \pi^*(a\,|\,x)|\,. \tag{6.1}$$

Then we have that the sequence defined by choosing d_0 arbitrarily and setting $d_{k+1} = F_d^{\pi_k} d_k$ satisfies $d_k \to d^{\pi^*}$ as $k \to \infty$.

Proof. Let $\varepsilon > 0$. Then let $K > 0$ be sufficiently large so that for all $k \geq K$, we have

$$\sup_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} |\pi_k(a|x) - \pi^*(a|x)| < \varepsilon\,.$$

We first demonstrate that closeness of π to π^* implies closeness in the sense of supremum norm of F_d^π to $F_d^{\pi^*}$:

$$
\begin{aligned}
\|F_d^\pi - F_d^{\pi^*}\|_\infty \leq \|\mathcal{R}^\pi - \mathcal{R}^{\pi^*}\|_\infty + \gamma \|\mathcal{P}^\pi - \mathcal{P}^{\pi^*}\|_\infty \\
&= \sup_{x \in \mathcal{X}} \left| \sum_{a \in \mathcal{A}} (\pi(a|x) - \pi^*(a|x)) \mathcal{R}_x^a \right| + \gamma \sup_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} \left| \mathcal{P}_x^\pi(x') - \mathcal{P}_x^{\pi^*}(x') \right| \\
&\leq \sup_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \left| \pi(a|x) - \pi^*(a|x) \right| R_{\text{max}} + \gamma \sup_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} \left| \sum_{a \in \mathcal{A}} \mathcal{P}_x^a(x') \pi(a|x) - \mathcal{P}_x^a(x') \pi^*(a|x) \right| \\
&\leq \sup_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} |\pi(a|x) - \pi^*(a|x)| R_{\text{max}} + \gamma \sup_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} \sum_{a \in \mathcal{A}} |\mathcal{P}_x^a(x') \pi(a|x) - \mathcal{P}_x^a(x') \pi^*(a|x)| \\
&= \sup_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} |\pi(a|x) - \pi^*(a|x)| R_{\text{max}} + \gamma \sup_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} \sum_{a \in \mathcal{A}} \mathcal{P}_x^a(x') |\pi(a|x) - \pi^*(a|x)| \\
&= (R_{\text{max}} + \gamma) \sup_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \left| \pi(a|x) - \pi^*(a|x) \right| .
\end{aligned}
$$

Hence, for all $k \geq K$, we have $\|F_d^{\pi_k} - F_d^{\pi^*}\|_\infty \leq (R_{\text{max}} + \gamma)\varepsilon$. We therefore have

$$
\begin{aligned}
d_{k+1} &= F_d^{\pi_k} d_k \\
\implies d_{k+1} - d^{\pi^*} &= F_d^{\pi_k} d_k - d^{\pi^*} \\
\implies d_{k+1} - d^{\pi^*} &= F_d^{\pi_k} d_k - F_d^{\pi_k} d^{\pi^*} + F_d^{\pi_k} d^{\pi^*} - d^{\pi^*} \\
\implies \|d_{k+1} - d^{\pi^*}\|_\infty &\leq \|F_d^{\pi_k} d_k - F_d^{\pi_k} d^{\pi^*}\|_\infty + \|F_d^{\pi_k} d^{\pi^*} - d^{\pi^*}\|_\infty \\
\implies \|d_{k+1} - d^{\pi^*}\|_\infty &\leq \gamma \|d_k - d^{\pi^*}\|_\infty + \|(F_d^{\pi_k} - F_d^{\pi^*}) d^{\pi^*}\|_\infty \\
\implies \|d_{k+1} - d^{\pi^*}\|_\infty &\leq \gamma \|d_k - d^{\pi^*}\|_\infty + (R_{\text{max}} + \gamma)\varepsilon \|d^{\pi^*}\|_\infty .
\end{aligned}
$$

Now letting $z_k = \|d_k - d^{\pi^*}\|_\infty$, this gives

$$
z_{k+1} \leq \gamma z_k + (R_{\text{max}} + \gamma)\varepsilon \|d^{\pi^*}\|_\infty .
$$

Taking limit superior on each side, we obtain

$$\limsup z_k \leq \gamma \limsup z_k + (R_{\text{max}} + \gamma)\varepsilon \|d^{\pi^*}\| .$$

Rearranging gives $\limsup z_k \leq (1-\gamma)^{-1}(R_{\text{max}} + \gamma)\varepsilon\|d^{\pi^*}\|$. Since $\varepsilon > 0$ was arbitrary, we have $\limsup z_k \leq 0$, but $(z_k)_{k\geq 0}$ is a non-negative sequence, hence $\lim z_k = 0$, as required.

□

Hence in order to prove that a distance d^π satisfies convergence under changing policies, it suffices to show that the operator F_d satisfies the Lipschitz condition (5.1). We now prove that this is indeed the case for the metrics considered thus far.

Theorem 6.2.3. *All state similarity metrics introduced in section 6.1 satisfy convergence under changing policies as described in Theorem 6.2.2.*

Proof. As remarked above, it remains to show that each of the distances satisfy the Lipschitz condition (5.1). Rather than prove this individually for each distance, we will use the result from section 6.1, that is, each of these metric's operator is actually the Bellman operator T^π for an auxiliary MDP $\tilde{M}$. Therefore, it is sufficient to prove that T^π satisfies (5.1). We can now see this as

$$\begin{aligned}
\|T^\pi - T^{\pi^*}\|_\infty &= \sup_{x\in\mathcal{X}} \left| \mathcal{R}^\pi_x + \gamma \sum_{x'\in\mathcal{X}} \mathcal{P}^\pi_x(x') - \left(\mathcal{R}^{\pi^*}_x + \gamma \sum_{x'\in\mathcal{X}} \mathcal{P}^{\pi^*}_x(x') \right) \right| \\
&\leq \sup_{x\in\mathcal{X}} \left\{ \left| \mathcal{R}^\pi_x - \mathcal{R}^{\pi^*}_x \right| + \gamma \left| \sum_{x'\in\mathcal{X}} \mathcal{P}^\pi_x(x') - \sum_{x'\in\mathcal{X}} \mathcal{P}^{\pi^*}_x(x') \right| \right\} \\
&\leq \sup_{x\in\mathcal{X}} \left\{ \sum_{a\in\mathcal{A}} |\mathcal{R}^a_x \, \pi(a\,|\,x) - \mathcal{R}^a_x \, \pi^*(a\,|\,x)| + \gamma \sum_{x'\in\mathcal{X}} \sum_{a\in\mathcal{A}} |\mathcal{P}^a_x(x')\pi(a\,|\,x) - \mathcal{P}^a_x(x')\pi^*(a\,|\,x)| \right\} \\
&= \sup_{x\in\mathcal{X}} \left\{ \sum_{a\in\mathcal{A}} |\mathcal{R}^a_x \, \pi(a\,|\,x) - \mathcal{R}^a_x \, \pi^*(a\,|\,x)| + \gamma \sum_{x'\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mathcal{P}^a_x(x') \, |\pi(a\,|\,x) - \pi^*(a\,|\,x)| \right\} \\
&\leq \sup_{x\in\mathcal{X}} \left\{ (2R_{\text{max}} + \gamma) \sum_{a\in\mathcal{A}} |\pi(a\,|\,x) - \pi^*(a\,|\,x)| \right\} \\
&= (2R_{\text{max}} + \gamma) \sup_{x\in\mathcal{X}} \left\{ \sum_{a\in\mathcal{A}} |\pi(a\,|\,x) - \pi^*(a\,|\,x)| \right\},
\end{aligned}$$

from which we can see that the desired condition holds with $C = 2R_{\text{max}} + \gamma$. $\square$

6.3 Discussion and future work

In this chapter, we discuss a framework of auxiliary Markov decision processes, whose value functions correspond to the state similarity metrics discussed in this thesis. This has produced novel distances, and illustrates the relationships between the various metrics.

One direction for future work is investigating the metrics produced for various couplings. In a given auxiliary setting, the MDP produced is uniquely determined by the choice of the transition coupling λ. In this chapter, we have exclusively considered the optimal and independent couplings, which produce metrics which use the Kantorovich metric and Łukaszyk–Karmowski distance respectively. These can be seen as two extremes, and one may find that more complex couplings may be able to interpolate between the two.

Chapter 7

Conclusion

In this thesis we have considered many aspects of state similarity metrics in reinforcement learning.

We firstly introduced the MICo distance, a state similarity metric inspired by bisimulation which can be computed efficiently from samples. We proved that it satisfied desirable properties subsection 3.2.3, and demonstrated that it can be learned through a loss in deep learning settings. We then present empirical results, showing that representation learning using the MICo distance is effective for large-scale settings.

Next, we turned from the problem of learning behavioural metrics on Markov decision processes to instead learning behavioural kernels on Markov decision processes. From these, we leveraged the equivalence of kernels and metrics of negative type to recover a new distance, the kernel similarity metric. We then proved its equivalence to the reduced MICo distance from the previous paragraph, which gave us new theoretical properties and insights.

Next, we take a distributional perspective on the previous chapters and literature, and consider what happens when deterministic reward assumptions are dropped. We introduced new behavioural metrics constructed using suitable modification to the original definitions, and showed that using the same update schemes in general reward settings we obtain these new metrics. We then proved various properties of these metrics, and related them to recent concepts in distributional reinforcement learning.

Finally, we present a unifying framework of auxiliary Markov decision processes, and how state similarity metrics can be extracted as the value functions of these MDPS. Through this perspective, all metrics considered thus far, as well as a number of others, were shown to arise as value functions for various MDPs. Moreover, we used this framework to prove that all of these metrics satisfy convergence under changing policies.

For each of these topics, we have provided fruitful areas for future research (refer to section 3.5, section 4.4, section 5.2, and section 6.3 respectively).

Bibliography

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

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. URL `http://dx.doi.org/10.2307/1990404`.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922. URL `http://eudml.org/doc/213289`.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, jun 2013.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2022. `http://www.distributional-rl.org`.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, Belmont, MA, 1996.

R. Blute, J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for labelled Markov processes. In *Proceedings of the Twelfth IEEE Symposium On Logic In Computer Science, Warsaw, Poland.*, 1997.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic Markov decision processes. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10069–10076. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6564.

Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G. Bellemare. Dopamine: A Research Framework for Deep Reinforcement Learning. 2018. URL http://arxiv.org/abs/1812.06110.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. Mico: Learning improved representations via sampling-based state similarity for Markov decision processes. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*, 2021. URL https://arxiv.org/abs/2106.08229.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Will Dabney, André Barreto, Mark Rowland, Robert Dadashi, John Quan, Marc G Bellemare, and David Silver. The value-improvement path: Towards better representations for reinforcement learning. *arXiv preprint arXiv:2006.02243*, 2020.

J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Metrics for labeled Markov systems. In *Proceedings of CONCUR99*, number 1664 in Lecture Notes in Computer Science. Springer-Verlag, 1999.

J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for labeled Markov processes. *Information and Computation*, 179(2):163–193, Dec 2002.

Josée Desharnais, Vineet Gupta, Radhakrishnan Jagadeesan, and Prakash Panangaden. A metric for labelled Markov processes. *Theoretical Computer Science*, 318(3):323–354, June 2004.

Regina C. Elandt. The folded normal distribution: Two methods of estimating parameters from moments. *Technometrics*, 3(4):551–562, 1961. doi: 10.1080/00401706.1961.10489975. URL `https://www.tandfonline.com/doi/abs/10.1080/00401706.1961.10489975`.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *UAI*, volume 4, pages 162–169, 2004.

Norm Ferns, Pablo Samuel Castro, Doina Precup, and Prakash Panangaden. Methods for computing state similarity in Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

Norman Ferns and Doina Precup. Bisimulation metrics are optimal value functions. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

György Pál Gehér, Tamás Titkos, and Dániel Virosztek. The isometry group of wasserstein spaces: the hilbertian case. *arXiv preprint arXiv:2102.02037*, 2021.

C. Gini. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. [Fasc. I.]*. Studi economico-giuridici pubblicati per cura della facoltà di Giurisprudenza della R. Università di Cagliari. Tipogr. di P. Cuppini, 1912. URL `https://books.google.ca/books?id=fqjaBPMxB9kC`.

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

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012a. URL `http://jmlr.org/papers/v13/gretton12a.html`.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012b. URL `https://proceedings.neurips.cc/paper/2012/file/dbe272bab69f8e13f14b405e038deb64-Paper.pdf`.

C. Guilbart. Produits scalaires sur l'espace des mesures. *Annales de l'I.H.P. Probabilités et statistiques*, 15(4):333–354, 1979.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Leonid V. Kantorovich and G. Sh. Rubinshtein. On a space of totally additive functions. *Vestn Lening. Univ.*, 13(7):52–59, 1958.

Mete Kemertas and Tristan Aumentado-Armstrong. Towards robust bisimulation metric learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Mete Kemertas and Allan Jepson. Trusted approximate policy iteration with bisimulation metrics. *arXiv preprint arXiv:2202.02881*, 2022.

Charline Le Lan, Stephen Tu, Adam Oberman, Rishabh Agarwal, and Marc G. Bellemare. On the generalization of representations in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

K. G. Larsen and A. Skou. Bisimulation through probablistic testing. *Information and Computation*, 94:1–28, 1991.

Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. On the effect of auxiliary tasks on representation dynamics. *CoRR*, abs/2102.13089, 2021. URL `https://arxiv.org/abs/2102.13089`.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew J. Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61: 523–562, 2018.

S. G. Matthews. Partial metric topology. *Annals of the New York Academy of Sciences*, 728 (1):183–197, 1994.

R. Milner. *A Calculus for Communicating Systems*, volume 92 of *Lecture Notes in Computer Science*. Springer-Verlag, 1980.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Prakash Panangaden. *Labelled Markov Processes*. Imperial College Press, GBR, 2009. ISBN 1848162871.

David Park. Title unknown. Slides for Bad Honnef Workshop on Semantics of Concurrency, 1981.

Bilal Piot, Matthieu Geist, and Olivier Pietquin. Difference of convex functions programming for reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

Mark Rowland, Marc Bellemare, Will Dabney, Remi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 29–37. PMLR, 09–11 Apr 2018. URL `https://proceedings.mlr.press/v84/rowland18a.html`.

Bernhard Schölkopf, Alexander J. Smola, and Francis Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018. ISBN 0262536579.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, illustrated edition edition, 2004. ISBN 0521813972. URL `http://www.amazon.com/Kernel-Methods-Pattern-Analysis-Shawe-Taylor/dp/0521813972`.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016. doi: 10.1038/nature16961.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, and Bernhard Schölkopf. Kernel choice and classifiability for rkhs embeddings of probability distributions. pages 1750–1758, 01 2009a.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert R. G. Lanckriet, and Bernhard Schölkopf. A note on integral probability metrics and ϕ-divergences. *CoRR*, abs/0901.2698, 2009b. URL `http://arxiv.org/abs/0901.2698`.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none):1550 – 1599, 2012. doi: 10.1214/12-EJS722. URL `https://doi.org/10.1214/12-EJS722`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL `http://incompleteideas.net/book/the-book-2nd.html`.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 5005–5014. PMLR, 2018.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4235–4246, 2020.

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

Shlomo Yitzhaki. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron - International Journal of Statistics*, LXI:285–316, 02 2003.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=-2FCwDKRREu`.

Szymon Łukaszyk. A new concept of probability metric and its applications in approximation of scattered data sets. *Computational Mechanics*, 33:299–304, 03 2004.