Bisimulation Metrics are Optimal Value Functions

Norm Ferns *

Département d'Informatique
École Normale Supérieure
45 rue d'Ulm, F-75230 Paris Cedex 05, France

Doina Precup †

School of Computer Science
McGill University
Montréal, Canada, H3A 0E7

Abstract

Bisimulation is a notion of behavioural equivalence on the states of a transition system. Its definition has been extended to Markov decision processes, where it can be used to aggregate states. A bisimulation metric is a quantitative analog of bisimulation that measures how similar states are from a the perspective of long-term behavior. Bisimulation metrics have been used to establish approximation bounds for state aggregation and other forms of value function approximation. In this paper, we prove that a bisimulation metric defined on the state space of a Markov decision process is the optimal value function of an optimal coupling of two copies of the original model. We prove the result in the general case of continuous state spaces. This result has important implications in understanding the complexity of computing such metrics, and opens up the possibility of more efficient computational methods.

1 INTRODUCTION

Markov decision processes (MDPs) are a popular mathematical model for sequential decision-making under uncertainty (Puterman, 1994; Sutton & Barto, 2012). Many standard solution methods are based on computing or learning the optimal value function, which reflects the expected return one can achieve in each state by choosing actions according to the optimal policy. In finite MDPs, the optimal value function is guaranteed to be unique, and has at least one deterministic optimal policy associated with it.

A major challenge is how to deal with large, possibly continuous, state spaces, known more colourfully as *the curse of dimensionality* or *the state-space explosion problem*. Briefly, the number of parameters necessary to represent the value function scales exponentially with the number of state variables. In response to this issue, a number of researchers have advocated the use of metrics, which can be used to determine similarity between states, and cluster them accordingly. Ideally, one would like such a clustering to reflect similarity among states in terms of the value function, which reflects the long-term cumulative reward.

In the formal verification community, a similar problem arises in the analysis of transition systems, in which one wants to establish long-term properties (e.g., the probability that the system may enter a faulty state, or that a certain trajectory would terminate). Many researchers advocate tackling such problems by using approximation metrics based on strong probabilistic bisimulation. Bisimulation is a conservative behavioural equivalence between states: states that are bisimilar will have the same long-term behaviour (Larsen & Skou, 1991; Givan et al., 2003). Corresponding metrics are useful in order to measure state similarity, and are used both to directly aggregate system states and more generally to assess the quality of an approximation. Abate (2012) surveys historical and more recent developments in this area.

In the context of MDPs, such metrics - henceforth known as *bisimulation metrics* - were developed in (Ferns et al., 2004; Ferns et al., 2005; Ferns et al., 2011) based on the work of Desharnais et al. (2002; 2001a) for a related Markov transition system. In (Ferns et al., 2006), the authors experimentally compared several methods for estimating these metrics on small finite MDPs, with a Monte Carlo approach outperforming the others. However, the analyses therein were limited, lacking, for example, any sample complexity results.

The purpose of this work is to strengthen and unify theoretical and practical results for bisimulation metrics on a given MDP by showing that they are in fact the optimal value functions of an optimal coupling of that MDP with itself (Theorem 3.3). We establish this result in the general setting of continuous-state MDPs. To our knowledge, this is an original result, which both improves our under-

* Norm Ferns' contribution was partially supported by the AbstractCell ANR-Chair of Excellence.

† Doina Precup's contribution was supported by NSERC.

standing of bisimulation metrics in general, and opens up avenues of attack for more efficient computation.

The paper is organized as follows. In Section 2, we provide a brief summary of MDPs, optimal control theory, and bisimulation metrics on continuous spaces. In Section 3, we use a measurable selection theorem to prove the main result, and in Section 4 we relate the results we present to existing work within the artificial intelligence and formal verification communities. Finally, in Section 5, we discuss the implications of our result and directions for future research.

2 BACKGROUND

Since we deal primarily with uncountably infinite state spaces, we must take into account the tension between imposing the right amount of structure on a space for general theoretical usefulness and imposing the right amount of structure for useful practical applications. For that reason, much of the work on Markov processes has been cast in the setting of Polish spaces. The introductory chapter of (Doberkat, 2007) contains a self-contained exposition of probabilities on Polish spaces in the context of computer science. A more comprehensive mathematical description can be found in (Srivastava, 2008). By contrast, a gentler introduction to probabilities on continuous spaces can be found in the first four chapters of (Folland, 1999). We refer the reader to these three sources for the basic mathematical definitions that we present throughout.

2.1 PROBABILITIES ON METRIC SPACES

A *Polish metric space* is a complete, separable metric space. A *Polish space* is a topological space that is homeomorphic to a Polish metric space. A *standard Borel space* is a measurable space that is Borel isomorphic to a Polish space.

If (X, τ) is a topological space, then $\mathbb{C}^b(X)$ is the set of continuous bounded real-valued functions on X. If $(X, \mathcal{B}_X)$ is a standard Borel space then we denote by $\mathbb{B}^b(X)$ the space of bounded measurable real-valued functions on X, and by $\mathbb{P}(X)$ the set of probability measures on X; note that the latter is also a standard Borel space (Giry, 1982).

2.2 DISCOUNTED MARKOV DECISION PROBLEMS

Let $(X, \mathcal{B}_X)$ and $(Y, \mathcal{B}_Y)$ be standard Borel spaces. A *Markov kernel* is a Borel measurable map from $(X, \mathcal{B}_X)$ to $(\mathbb{P}(Y), \mathcal{B}_{\mathbb{P}(Y)})$. Equivalently, K is a Markov kernel from X to Y iff $K(x)$ is a probability measure on $(Y, \mathcal{B}_Y)$ for each $x \in X$, and $x \mapsto K(x)(B)$ is a measurable map for each $B \in \mathcal{B}_Y$.

Remark 1. *The use of the term* kernel *here should not be confused with the usual meaning, that being the set of points in the domain of a real-valued function that send the function to* 0*. We will refer to kernel in both senses throughout, with the meaning clear from the context.*

We will denote the set of all Markov kernels from X to Y by $[\![X \to \mathbb{P}(Y)]\!]$ and simply write "K is a Markov kernel on X" when it is implicitly assumed that $Y = X$. If I is an index set and $K = (K_i)_{i \in I}$ is an I-indexed collection of Markov kernels on X, we will say that "K is a labelled Markov kernel on X". Such kernels play the role of transition relations in stochastic transition systems with continuous state spaces.

A *Markov decision process (MDP)* is a tuple $(S, \mathcal{B}_S, A, (P_a)_{a \in A}, r)$, where $(S, \mathcal{B}_S)$ is a standard Borel space, A is a finite set of actions, $r : A \times S \to \mathbb{R}$ is a bounded measurable reward function, and for $a \in A$, P_a is a Markov kernel on S.

For each $a \in A$, we denote by $r_a : S \to \mathbb{R}$ the function defined by $r_a(s) = r(a, s)$, and for each $a \in A$ and $s \in S$. We use functional notation for integration with respect to $P_a(s)$, i.e. the integral of $f \in \mathbb{B}^b(S)$ with respect to $P_a(s)$ will be written as $P_a(s)(f)$.

A Markov decision process along with an optimality criterion is known as a *Markov decision problem.* In this work, we focus on Markov decision problems with the *expected total discounted reward optimality criterion*, which we now briefly describe based on (Hernández-Lerma & Lasserre, 1996) and especially Section 8.3 of (Hernández-Lerma & Lasserre, 1999). We rely on these sources instead of others who may be more familiar to the AI audience because they treat the infinitely uncountable state space setting. We direct the reader to these sources for full details.

Fix an MDP $\mathcal{M} = (S, \mathcal{B}_S, A, (P_a)_{a \in A}, r)$ and a discount factor $\gamma \in (0, 1)$. Let $t \in \mathbb{N}$. Then H_t, the family of *histories up to time* t, is defined by $H_0 = S$ and $H_{t+1} = H_t \times (A \times S)$ for $t \in \mathbb{N}$. An element $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t) \in H_t$ is called a *t-history*. A *randomized control policy* is a sequence of Markov kernels $\pi = (\pi_t)_{t \in \mathbb{N}}$ such that $\pi_t \in [\![H_t \to \mathbb{P}(A)]\!]$ for all $t \in \mathbb{N}$. The set of all policies is denoted by Π. A policy $\pi = (\pi_t)_{t \in \mathbb{N}}$ is said to be a *randomized stationary policy* if there exists a Markov kernel $\varphi \in [\![S \to \mathbb{P}(A)]\!]$ such that $\pi_t(h_t) = \varphi(s_t)$ for all $h_t \in H_t, t \in \mathbb{N}$, and a *deterministic stationary policy* if there exists a measurable selector f for $S \times A$ such that $\pi_t(h_t)$ is the Dirac measure at the point $f(s_t) \in A$ for all $h_t \in H_t, t \in \mathbb{N}$. We denote the sets of randomized stationary policies and deterministic stationary policies by Π_{RS} and Π_{DS} respectively and note that $\Pi_{DS} \subseteq \Pi_{RS} \subseteq \Pi$.

Let $\pi \in \Pi$ be a policy on $\mathcal{M}$. The *γ-discounted value function for* π, $V_\gamma(\pi)$, is defined by $V_\gamma(\pi)(s) =$

$\mathbb{E}_s^\pi[\sum_{t=0}^\infty \gamma^t r(a_t, x_t)]$ for all $s \in S$, where $\mathbb{E}_s^\pi$ is the expectation taken with respect to the system dynamics when starting in state s and following policy π. The goal of this Markov decision problem is to find a policy whose value function dominates all others. Toward that end, one defines the γ-discounted optimal value function, V_γ^*, by $V_\gamma^*(s) = \sup_{\pi\in\Pi} V_\gamma(\pi)(s)$ for all $s \in S$, and the *Bellman operator* with respect to γ, $T_\gamma : \mathbb{B}^b(S) \to \mathbb{B}^b(S)$, by $T_\gamma(v)(s) = \max_{a\in A}[r_a(s) + \gamma \cdot P_a(s)(v)]$ for all $s \in S$. The following can be found within Theorem 8.3.6 and its preceding remarks in (Hernández-Lerma & Lasserre, 1999).

Theorem 2.1 (Value Iteration)**.** *Define $(v_n)_{n\in\mathbb{N}} \subseteq \mathbb{B}^b(S)$ by $v_0(s) = 0$ for all $s \in S$ and $v_{n+1} = T_\gamma(v_n)$ for all $n \in \mathbb{N}$. Then the optimal value function V_γ^* is the unique solution in $\mathbb{B}^b(S)$ to the* Bellman optimality equation *$v = T_\gamma(v)$; $(v_n)_{n\in\mathbb{N}}$ converges uniformly to V_γ^* with $\|v_n - V_\gamma^*\| \leq \gamma^{-n}(1-\gamma)$ for all $n \in \mathbb{N}$ and where $\|\cdot\|$ is the uniform norm; and there exists a deterministic stationary optimal policy $\pi^* \in \Pi_{DS}$ such that $V_\gamma^* = V_\gamma(\pi^*)$.*

We note that although Theorem 2.1 implies it is sufficient to search Π_{DS} for an optimal policy, in practice it is often useful to work with the larger class Π_{RS}. On the other hand, for more general theoretical considerations, e.g. other optimality criteria, we may need to consider all of Π.

2.3 BISIMULATION

We present bisimulation for MDPs as outlined in (Ferns et al., 2011).

Given an equivalence relation R on a measurable space (S, Σ), a subset X of S is said to be *R-closed* if X is a union of R-equivalence classes. We write $\Sigma(R)$ for the set of those Σ-measurable sets that are also R-closed.

Let $(S, \mathcal{B}_S, A, (P_a)_{a\in A}, r)$ be an MDP. An equivalence relation R on S is a *bisimulation relation* if it satisfies $sRs' \iff$ for every $a \in A$, $r_a(s) = r_a(s')$ and for every $X \in \Sigma(R)$, $P_a(s)(X) = P_a(s')(X)$. *Bisimilarity* is the largest of the bisimulation relations.

2.4 THE KANTOROVICH METRIC

In order to define bisimulation metrics for MDPs, we first need to recall the definition and properties of the Kantorovich metric between distributions, which can be found in (Villani, 2003).

Definition 1 (Kantorovich Metric)**.** *Let S be a Polish space, h a bounded pseudometric on S that is lower semi-continuous on $S \times S$ with respect to the product topology, and let $Lip(h)$ be the set of all $f \in \mathbb{B}^b(S)$ that satisfy the Lipschitz condition $f(x) - f(y) \leq h(x, y)$ for every $x, y \in S$. Let $P, Q \in \mathbb{P}(S)$. Then the* Kantorovich metric *$\mathcal{K}(h)$ is defined by $\mathcal{K}(h)(P, Q) = \sup_{f\in Lip(h)}(P(f) - Q(f))$.*

The Kantorovich metric is an infinite linear program and has a dual described in terms of couplings of probability measures.

Definition 2 (Coupling)**.** *Let $(X, \mathcal{B}_X)$ and $(Y, \mathcal{B}_Y)$ be standard Borel spaces, and let $(X \times Y, \mathcal{B}_X \otimes \mathcal{B}_Y)$ be the product space. Let $\mu \in \mathbb{P}(X)$, $\nu \in \mathbb{P}(Y)$, and $\lambda \in \mathbb{P}(X \times Y)$. Then λ is a* coupling *of μ and ν if and only if its marginals on X and Y are μ and ν, respectively. We denote the set of all couplings of μ and ν by $\Lambda(\mu, \nu)$, i.e., $\lambda \in \Lambda(\mu, \nu) \iff \lambda(E \times Y) = \mu(E)$ and $\lambda(X \times F) = \nu(F)$ for all $E \in \mathcal{B}_X, F \in \mathcal{B}_Y$.*

The following is found within Section 1.1.1 of (Villani, 2003).

Lemma 2.2. *Let X and Y be Polish spaces and let μ and ν belong to $\mathbb{P}(X)$ and $\mathbb{P}(Y)$, respectively. Then $\lambda \in \Lambda(\mu, \nu)$ if and only if for every $(\varphi, \psi) \in \mathbb{C}^b(X) \times \mathbb{C}^b(Y)$*

$$\int_{X\times Y} [\varphi(x) + \psi(y)]\lambda(dx, dy) = \int_X \varphi(x)\mu(dx) + \int_Y \psi(y)\nu(dy).$$

In Section 3, we'll make use of the following simple lemma.

Lemma 2.3. *Let X and Y be Polish spaces and let μ and ν belong to $\mathbb{P}(X)$ and $\mathbb{P}(Y)$, respectively. Then $\Lambda(\mu, \nu)$ is a closed subset of $\mathbb{P}(X \times Y)$.*

Proof. Let $(\lambda_n)_{n\in\mathbb{N}} \subseteq \Lambda(\mu, \nu)$ be a sequence converging to some $\lambda \in \mathbb{P}(X \times Y)$ in the weak topology. Let $(\varphi, \psi) \in \mathbb{C}^b(X) \times \mathbb{C}^b(Y)$. Then

$$\begin{aligned}
&\int_{X\times Y} [\varphi(x) + \psi(y)]\lambda(dx, dy) \\
&= \lim_{n\to\infty} \left(\int_{X\times Y} [\varphi(x) + \psi(y)]\lambda_n(dx, dy) \right) \\
&= \lim_{n\to\infty} \left(\int_X \varphi(x)\mu(dx) + \int_Y \psi(y)\nu(dy) \right) \\
&= \int_X \varphi(x)\mu(dx) + \int_Y \psi(y)\nu(dy).
\end{aligned}$$

Here we have used the definition of weak convergence, as well as Lemma 2.2 for each λ_n. It follows from the same lemma that $\lambda \in \Lambda(\mu, \nu)$. □

The following can be found in Theorem 1.3 and the proof of Theorem 1.14 in (Villani, 2003).

Theorem 2.4 (Kantorovich-Rubinstein Duality Theorem)**.** *Assume the conditions of Definition 1. Then $\mathcal{K}(h)(P, Q)$ is equal to*

$$\sup_{f\in Lip(h,\mathbb{C}^b(S))} (P(f) - Q(f)) = \inf_{\lambda\in\Lambda(P,Q)} \lambda(h)$$

where $Lip(h, \mathbb{C}^b(S))$ denotes functions on S that are continuous and bounded, 1-Lipschitz with respect to h, and have range $[0, \|h\|]$. Moreover, the supremum and infimum are attained.

2.5 BISIMULATION METRICS

The following can be found in Theorem 3.12 of (Ferns et al., 2011) and Corollary 3 of (Ferns et al., 2014).

Theorem 2.5. *Let $\mathcal{M} = (S, \mathcal{B}_S, A, (P_a)_{a\in A}, r)$ be an MDP and let $c \in (0, 1)$ be a discount factor. Assume that the image of r is contained in $[0, 1]$. Then there exists a Polish topology τ generating $\mathcal{B}_S$ such that for all $a \in A$, r_a is continuous with respect to τ and P_a is weakly continuous with respect to τ. Define $\theta^c : A \times (S \times S) \to [0, 1]$ by $\theta^c_a(x, y) = (1 - c)|r_a(x) - r_a(y)|$. Furthermore, let $\mathfrak{lsc}_\mathfrak{m}$ be the set of bounded pseudometrics on S that are lower semicontinuous on $S \times S$ endowed with the product topology induced by τ. Define $F_c : \mathfrak{lsc}_\mathfrak{m} \to \mathfrak{lsc}_\mathfrak{m}$ by setting $F_c(\rho)(s, s')$ equal to*

$$\max_{a\in A}\left[\theta^c_a(s, s') + c \cdot \mathcal{K}(\rho)(P_a(s), P_a(s'))\right]$$

Then F_c has a unique 1-bounded fixed-point pseudometric $\rho^*_c \in \mathbb{C}^b(S \times S)$ whose kernel is bisimilarity.

We call such a metric a *bisimilarity metric*, and more generally a *bisimulation metric* if its kernel is a bisimulation relation (but not necessarily the largest). The following result, Theorem 3.20 in (Ferns et al., 2011), relates the optimal values of states to their similarity as measured by the bisimulation metric.

Theorem 2.6. *Assume the setup and result of Theorem 2.5. Let $\gamma \in (0, c]$ be a reward discount factor and let V^*_γ be the optimal value function defined in Theorem 2.1. Then V^*_γ is Lipschitz continuous with respect to ρ^*_c with Lipschitz constant $(1 - c)^{-1}$, i.e., for all $s, s' \in S$,*

$$|V^*_\gamma(s) - V^*_\gamma(s')| \leq (1 - c)^{-1}\rho^*_c(s, s')$$

Since bisimulation is a behavioural equivalence, this result implies that the closer two states are in bisimilarity distance, the more likely they are to share the same optimal actions, and hence optimal policies, for achieving the same optimal values.

3 A BISIMULATION VALUE FUNCTION

Let $\mathcal{M} = (S, \mathcal{B}_S, A, (P_a)_{a\in A}, r)$ be an MDP with the image of r contained in $[0, 1]$ and let $c \in (0, 1)$ be a discount factor. The goal of this section is to show that the bisimilarity metric ρ^*_c given by Theorem 2.5 can be expressed as the optimal value function of some MDP. In order to do so, let us first extend the definition of a coupling of two probability measures to a coupling of two labelled Markov kernels in the obvious way.

Definition 3. *Let $(X, \mathcal{B}_X)$ and $(Y, \mathcal{B}_Y)$ be standard Borel spaces, and let $(X \times Y, \mathcal{B}_X \otimes \mathcal{B}_Y)$ be the product space. Let I be an index set, and let $K = (K_i)_{i\in I}$, $L = (L_i)_{i\in I}$, and $M = (M_i)_{i\in I}$ be labelled Markov kernels on X, Y, and $X \times Y$, respectively. Then M is a* coupling *of K and L if and only if for each $i \in I$, $x \in X$, and $y \in Y$, $M_i(x, y)$ is a coupling of $K_i(x)$ and $L_i(y)$ in the sense of Definition 2. We denote the set of all couplings of K and L by $\Lambda(K, L)$.*

Recall that ρ^*_c is the unique solution to the fixed-point equation

$$\rho^*_c(x, y) = \max_{a\in A}[\theta^c_a(s, s') + c \cdot \mathcal{K}(\rho^*_c)(P_a(x), P_a(y))].$$

Here is the crucial fact: Theorem 2.4 remarkably not only provides a statement of duality for each Kantorovich linear program $\mathcal{K}(\rho^*_c)(P_a(x), P_a(y))$, but guarantees the existence of minimizers in the minimization linear program as well. Therefore, for every $a \in A$ and $x, y \in S$ there exists $\lambda_{axy} \in \Lambda(P_a(x), P_a(y))$ such that $\mathcal{K}(\rho^*_c)(P_a(x), P_a(y)) = \lambda_{axy}(\rho^*_c)$. Suppose for every $a \in A$ the map from $S \times S$ to $\mathbb{P}(S \times S)$ that sends (x, y) to λ_{axy} is measurable. Then ρ^*_c would satisfy the Bellman optimality equation defined in Theorem 2.1 for the optimal value function with discount factor c for the MDP $(S \times S, \mathcal{B}_{S\times S}, A, (\lambda_a)_{a\in A}, \theta^c)$ where $\lambda_a(x, y) = \lambda_{axy}$. Remark that such a $\lambda = (\lambda_a)_{a\in A}$ is a coupling of P with P, where $P = (P_a)_{a\in A}$. Thus, if we can find a measurable way of selecting the minimizers amongst all the Kantorovich linear programs appearing in the definition of ρ^*_c, we will have shown that the bisimilarity metric is actually the optimal value function of an MDP whose labelled Markov kernel is a coupling of two copies of the labelled Markov kernel of the original model. This is our goal.

3.1 MEASURABLE SELECTORS AND SECTIONS

The following results can be found in Section 1.4 of (Doberkat, 2007) and Section 5 of (Srivastava, 2008).

Let X and Y be sets. A *multifunction* from X to Y is a set-valued map $\mathcal{R} : X \to 2^Y$ such that for all $x \in X$, $\mathcal{R}(x)$ is a nonempty subset of Y. A multifunction $\mathcal{R}$ from X to Y can equivalently be viewed as a relation between X and Y.

Given a multifunction $\mathcal{R}$ between measurable spaces X and Y, one usually seeks to measurably select a member of $\mathcal{R}(x)$ for each $x \in X$. Here, we recount one way of doing so.

Let $\mathcal{R}$ be a multifunction from X to Y, and let $G \subseteq Y$. The *weak inverse* of G with respect to $\mathcal{R}$ is the set $\exists\mathcal{R}(G) = \{x \in X \mid \exists y \in G \text{ such that } (x, y) \in \mathcal{R}\} = \{x \in X \mid \mathcal{R}(x) \cap G \neq \emptyset\}$. The importance of the weak inverse lies in trying to utilise the property of measurability for a multifunction $\mathcal{R}$. A measurable function requires the preimage of every measurable set to be measurable. Here, we need only consider the preimages of compact sets.

Suppose $\mathcal{R}$ is a multifunction between a measurable space X and a Polish space Y. Let $G \subseteq Y$. If $\exists\mathcal{R}(G)$ is measurable whenever G is compact then $\mathcal{R}$ is called a *$\mathcal{C}$-measurable relation* on $X \times Y$.

Assume X is a measurable space, Y is Polish, $\mathcal{R}$ is a multifunction from X to Y and for each $x \in X$, $\mathcal{R}(x)$ is a non-empty closed subset of Y. Then a measurable map $f : X \to Y$ is called a *measurable selector* for $\mathcal{R}$ if and only if $f(x) \in \mathcal{R}(x)$ for all $x \in X$.

The following measurable selection result can be found as Proposition 1.57 and Proposition 1.58 in (Doberkat, 2007).

Proposition 3.1. *Assume that X is a measurable space, Y is a Polish space, and R is a $\mathcal{C}$-measurable relation on $X \times Y$. Then there exists a measurable selector f for $\mathcal{R}$.*

Finally, the following appears in Proposition 2.34 of (Folland, 1999), and will be used in conjunction with the preceding measurable selection theorem to establish our main result.

Proposition 3.2. *Suppose that $(X, \mathcal{B}_X)$, $(Y, \mathcal{B}_Y)$, and $(Z, \mathcal{B}_Z)$ are measurable spaces and that $f : X \times Y \to Z$ is a product-measurable function. Let $x \in X$. Define the X-*section of f at x, *$f_x : Y \to Z$, to be the function defined by $f_x(y) = f(x, y)$ for all $y \in Y$. Then f_x is measurable.*

3.2 BISIMILARITY AS A VALUE FUNCTION

Theorem 3.3. *Let us assume the setup and result of Theorem 2.5. Let $K = (K_a)_{a\in A} \in \Lambda(P, P)$, where $P = (P_a)_{a\in A}$. Define the coupling of $\mathcal{M}$ with itself through K to be the MDP $\mathcal{M}(K) = (S \times S, \mathcal{B}_{S\times S}, A, (K_a)_{a\in A}, \theta^c)$. Let $V_c^*(K)$ denote its optimal value function with respect to c, as defined in Theorem 2.1. Then there exists a $K^* \in \Lambda(P, P)$ such that $\rho_c^* = V_c^*(K^*) = \min_{K\in\Lambda(P,P)} V_c^*(K)$.*

In order to prove Theorem 3.3, we will need to make use of the following result, which can be found within the proof of Lemma 3.14 in (Ferns et al., 2011).

Lemma 3.4. *Assume the setup and result of Theorem 2.5. Then for each $a \in A$, the map sending (s, s') to $\mathcal{K}(\rho_c^*)(P_a(s), P_a(s'))$ is continuous on $S \times S$.*

Proof of Theorem 3.3. In order to prove the existence of K^*, we follow the method of part 3 of the proof of Lemma 4.9 in (Doberkat, 2007). First, however, we appeal to Theorem 2.5 to assert the existence of a Polish topology τ on S making each r_a continuous and each P_a weakly continuous for all $a \in A$. Let $X = A \times S \times S$ and $Y = \mathbb{P}(S \times S)$. The set A is a Polish space since it is finite, and X is a Polish space since it is a finite product of Polish spaces. Additionally, Y is also a Polish space (Giry, 1982). Define $\mathcal{R} : X \to 2^Y$ by setting $\mathcal{R}(a, x, y)$ to be the set of all $\lambda \in \Lambda(P_a(x), P_a(y))$ such that $\mathcal{K}(\rho_c^*)(P_a(x), P_a(y)) = \lambda(\rho_c^*)$. Theorem 2.4 implies that each $\mathcal{R}(a, x, y)$ is non-empty. Suppose $(\lambda_n)_{n\in\mathbb{N}} \subseteq \mathcal{R}(a, x, y)$ converges to $\lambda \in Y$. By Lemma 2.3, $\lambda \in \Lambda(P_a(x), P_a(y))$. Theorem 2.5 implies that $\rho_c^* \in \mathbb{C}^b(S \times S)$, so that by weak convergence $\lambda(\rho_c^*) = \lim_{n\to\infty} \lambda_n(\rho_c^*) = \mathcal{K}(\rho_c^*)(P_a(x), P_a(y))$. Therefore, $\lambda \in \mathcal{R}(a, x, y)$, i.e. $\mathcal{R}(a, x, y)$ is closed.

Next, let $G \subseteq Y$ be compact, hence, closed. We will show that $\exists\mathcal{R}(G)$ is closed, and hence measurable. Let $(a_n, x_n, y_n)_{n\in\mathbb{N}} \subseteq \exists\mathcal{R}(G)$ be a sequence converging to some $(a, x, y) \in X$. So there exists $(\lambda_n)_{n\in\mathbb{N}} \subseteq G$ such that $\lambda_n \in \mathcal{R}(a_n, x_n, y_n)$ for all $n \in \mathbb{N}$. Since G is compact, it is also sequentially compact and therefore there exists a subsequence $(\lambda_{k_n})_{n\in\mathbb{N}}$ converging to some $\lambda \in G$. Remark that since A is finite, the sequence $(a_n)_{n\in\mathbb{N}}$ is eventually constant, i.e. there exists $N \in \mathbb{N}$ such that $(a_n, x_n, y_n) = (a, x_n, y_n)$ for all $n \geq N$. Let $(\varphi, \psi) \in \mathbb{C}^b(S) \times \mathbb{C}^b(S)$. Then

$$\begin{aligned}
&\int_{S\times S} [\varphi(s) + \psi(s')]\lambda(ds, ds') \\
&= \lim_{n\to\infty} \left(\int_{S\times S} [\varphi(s) + \psi(s')]\lambda_{k_n}(ds, ds') \right) \\
&= \lim_{n\to\infty} \Big(\int_S \varphi(s) P_{a_{k_n}}(x_{k_n})(ds) \\
&\qquad\qquad + \int_S \psi(s') P_{a_{k_n}}(y_{k_n})(ds') \Big) \\
&= \lim_{n\to\infty} \Big(\int_S \varphi(s) P_a(x_{k_n})(ds) \\
&\qquad\qquad + \int_S \psi(s') P_a(y_{k_n})(ds') \Big) \\
&= \int_S \varphi(s) P_a(x)(ds) + \int_S \psi(y) P_a(y)(ds')
\end{aligned}$$

so that $\lambda \in \Lambda(P_a(x), P_a(y))$. Here we have used the weak convergence of $(\lambda_{k_n})_{n\in\mathbb{N}}$ to λ, $(P_a(x_{k_n}))_{n\in\mathbb{N}}$ to $P_a(x)$, and of $(P_a(y_{k_n}))_{n\in\mathbb{N}}$ to $P_a(y)$, and the repeated use of Lemma 2.2. Moreover, by weak convergence and Lemma 3.4

$$\begin{aligned}
\lambda(\rho_c^*) &= \lim_{n\to\infty} \lambda_{k_n}(\rho_c^*) \\
&= \lim_{n\to\infty} \mathcal{K}(\rho_c^*)(P_a(x_{k_n}), P_a(y_{k_n})) \\
&= \mathcal{K}(\rho_c^*)(P_a(x), P_a(y)),
\end{aligned}$$

whence it follows that $\lambda \in \mathcal{R}(a, x, y)$. Therefore, $\mathcal{R}(a, x, y) \cap G \neq \emptyset$, $(a, x, y) \in \exists\mathcal{R}(G)$, and $\exists\mathcal{R}(G)$ is closed and hence measurable. By definition, $\mathcal{R}$ is a $\mathcal{C}$-measurable relation on $X \times Y$. Applying Proposition 3.1 there exists a measurable selector $f : X \to Y$ for $\mathcal{R}$. Finally, set $K^* = (K_a^*)_{a\in A}$ where $K_a^*(x, y) = f(a, x, y) \in \mathcal{R}(a, x, y)$ for all $a \in A$, $x, y \in S$. For each $a \in A$, K_a^* is simply the A-section of f at a, so that by Proposition 3.2, each $K_a^* \in [\![S \times S \to \mathbb{P}(S \times S)]\!]$. Therefore, $\rho_c^* = V_c^*(K^*)$, the optimal value function for $\mathcal{M}(K^*)$.

Clearly $\inf_{K\in\Lambda(P,P)} V_c^*(K) \leq V_c^*(K^*)$. To establish the reverse inequality, let $K = (K_a)_{a\in A} \in \Lambda(P, P)$. Then for

any $a \in A$ and $x, y \in S$,

$$\begin{aligned}
&\theta_a^c(x,y) + c \cdot K_a^*(x,y)(\rho_c^*) \\
&= \theta_a^c(x,y) + c \cdot \mathcal{K}(\rho_c^*)(P_a(x), P_a(y)) \\
&= \theta_a^c(x,y) + c \cdot \inf_{\lambda \in \Lambda(P_a(x), P_a(y))} \lambda(\rho_c^*) \\
&\leq \theta_a^c(x,y) + c \cdot K_a(x,y)(\rho_c^*).
\end{aligned}$$

By taking the maximum over all $a \in A$ and noting that the result holds for all $x, y \in S$, we then obtain $\rho_c^* \leq T_c(K)(\rho_c^*)$, where $T_c(K)$ is the Bellman optimality operator for the MDP $\mathcal{M}(K)$. Therefore, it follows that $V_c^*(K^*) \leq V_c^*(K)$ for any $K \in \Lambda(P,P)$, and finally that $V_c^*(K^*) \leq \inf_{K \in \Lambda(P,P)} V_c^*(K)$. □

Thus, we can interpret every discounted bisimulation metric as the optimal value function of some MDP; moreover, that MDP is optimal in the sense that it is the best coupling of the transition structure of the original MDP with itself when one seeks to minimize the expected total discounted *absolute difference* in rewards coming from the original model.

An immediate consequence is that we can now interpret the topology of convergence with respect to a bisimulation metric in terms of MDP optimality criteria. Conversely, when examining behavioural equivalence for the state space of a given MDP it no longer suffices to consider the structural model alone; one *must* take into account the full Markov decision problem, i.e. its intended use by means of an optimality criterion. This is yet another advantage of the pseudometric approach over that of exact equivalences. We discuss this further in Section **??**.

Practical implications are less immediate, but no less important, particularly in regard to determining what might be effective in attempting to calculate or estimate the distances. Consider a finite MDP. If we adjoin one new absorbing state with no immediate rewards then it is not hard to show that the bisimulation distance from that state to another state is the optimal value of the latter state. So computing a bisimulation metric is at least as hard as computing an optimal value function. On the other hand, we have just shown that computing a bisimulation metric amounts to computing an optimal value function - albeit, with the caveat that this amounts to a search over possibly infinitely many couplings. If one could restrict this search to polynomially many couplings, then it would follow that computing bisimulation metrics and computing optimal value functions belong to the same polynomial-time complexity class - and we conjecture that this is so. At first glance, this is a disappointing result; if one followed the naive approach, one would be attempting to solve for an optimal value function by solving for another optimal value function over a quadratically larger MDP. However, computing a bisimulation metric is of interest in its own right; the value function formulation allows for state-of-the-art reinforcement learning techniques (Sutton & Barto, 2012; Pazis & Parr, 2013) to be applied in its computation while at the same time informing us of what methods are unlikely to work well in practice for truly large systems.

A better practical approach would be to find more easily computable similarity metrics that are related in some meaningful way to the bisimulation metric. In that case, one would have a practical similarity measure with the theoretical guarantees given by bisimulation, as in Theorem 2.6. Our value function formulation permits a very natural way to do this, through the use of couplings.

Definition 4. *Let $f : X \times X \to [0,\infty)$ be a function on a set X. Define the function $\varrho(f) : X \times X \to [0,\infty)$ by $\varrho(f)(x,y) = \inf\{\sum_{j=1}^{m} \omega(f)(a_{j-1}, a_j)\}$, where $\omega(f)(u,v) = \min\{f(u,v), f(v,u)\}$ and the infimum is taken over all $m \in \mathbb{N}$ and $(a_j)_{j=0}^m \subseteq X$ such that $a_0 = x$ and $a_m = y$. Then $\varrho(f)$ is the largest pseudometric less than f.*

Notice that if X is finite and f is computable then the problem of computing $\varrho(f)$ is the *All-Pairs Shortest Paths* problem.

Corollary 3.5. *Assume the setup and result of Theorem 3.3. For $\pi \in \Pi$ defined over $S \times S$ we let $V_c(K)(\pi)$ denote the value function of $\mathcal{M}(K)$ with respect to π and c and we let $V_c^*(K)$ denote its optimal value function with respect to c, as defined in Theorem 2.1. Then*

1. $\forall K \in \Lambda(P,P)$, $\rho_c^* \leq \varrho(V_c^*(K)) \leq V_c^*(K)$.
2. $\forall \pi \in \Pi$, $\varrho(V_c(K^*)(\pi)) \leq V_c(K^*)(\pi) \leq \rho_c^*$.

Corollary 3.5 allows us to easily bound the bisimulation metric from above for any coupling $K \in \Lambda(P,P)$. For example, the product coupling $P \otimes P$ defined in the obvious way (and assuming measurability) by $(P \otimes P)_a(x,y) = P_a(x) \otimes P_a(y)$ should provide a trivial upper bound. Corollary 3.5 also provides a lower bound but only in the case where we know the optimal coupling K^* beforehand. Potentially more interesting is the case where we combine the two, i.e. for an arbitrary coupling $K \in \Lambda(P,P)$ and an arbitrary policy $\pi \in \Pi$ defined on $\mathcal{M}(K)$, does the equivalence induced by $\varrho(V_c(K)(\pi))$ lead to something that is more easily computable but that still provides good theoretical guarantees?

4 RELATED WORK

This work lies at the intersection of artificial intelligence and formal verification, and owes much to both. The concept of bisimulation has been in use within the uncertainty in artificial intelligence community for some time now. Indirectly in (Boutilier et al., 2000) and directly in (Givan et al., 2003), the notion of bisimulation had

been transferred from the theory of concurrent processes to MDP model minimization and the reinforcement learning paradigm. These papers work directly with factored or structured representations, which is an advantage over our approach for problems where such structure in the environment exists and is known explicitly. On the other hand, they deal only with discrete MDPs, exhibit the brittleness inherent in using exact equivalences for numerical systems, and lack theoretical guarantees on the size of a fully minimal system. In earlier work, Dean et al. (1997) actually consider an approximate version of bisimulation. For a small positive parameter ε they consider equivalence relations satisfying the property that immediate rewards and stochastic transitions to equivalence classes differ by at most ε. However, the disadvantages already mentioned still apply. More generally, (Li et al., 2006) provide a comprehensive survey and classification of various state abstractions for finite MDPs, including methods based on bisimulation (such as our bisimulation metrics).

Bisimulation metrics have been more extensively studied in the formal verification community. In that setting, the work closest in spirit to our own is (Chen et al., 2012), wherein the authors investigate the complexity of computing bisimilarity and metric bisimilarity for labelled Markov chains. In particular, Theorem 8 in that work relates an undiscounted bisimulation metric to optimal couplings of a given labelled Markov chain. Aside from considering only finite state systems, they allow for states to have differing sets of permissible actions but omit the reward parameter; hence, their work lies outside of the optimal control theory framework on which we focus.

Abate (2012) surveys various approximation metrics for probabilistic bisimulation over Markov processes with general state and action spaces, though here too Markov reward processes are mostly neglected. The author does conclude that a bridge needs to be made between techniques based on computing distances between Markov kernels and techniques based on sampling trajectories from processes under consideration; we believe the current work can help provide that bridge.

A very promising approach appears in (Desharnais et al., 2013) where the authors propose a general algorithm for estimating divergences, distance functions that may fail to satisfy the symmetry and triangle inequality axioms of a pseudometric. They consider divergences that generalize equivalences on probabilistic systems based on tests and observations. In particular, they define a new family of testable equivalences called *k-moment equivalences*; 1-moment equivalence is trace equivalence, as k grows larger k-moment equivalence becomes finer, and all k-moment equivalences as well as their limit equivalence are strictly weaker than bisimilarity. The exciting feature of their work is that the algorithm for estimating a divergence corresponding to a fixed equivalence is based on defining an MDP whose optimal value function is that divergence, and then using reinforcement learning techniques (Sutton & Barto, 2012) to solve for the optimal value function. While conceptually similar in spirit to our value function representation of bisimulation metrics, this approach differs significantly in how the MDP representing the metric is defined.

5 CONCLUSIONS AND FUTURE WORK

We have shown that the bisimulation metric defined in (Ferns et al., 2011) for an MDP is actually the optimal value function of an optimal coupling of the MDP with itself. This latter formulation is perhaps a more natural conception of distance for MDPs. In any case, all theoretical and practical results from optimal control theory concerning optimal value functions for MDPs can be carried over (based on this result)to the study of bisimulation metrics.

Perhaps the most intriguing implication of Theorem 3.3 is that examining other optimal control theory criteria may lead to different classes of bisimulation metrics perhaps better suited to those optimality tasks. Consider the undiscounted case. What does ρ_c^* represent when c tends to 1? We could set $c = 1$ in Theorem 2.5, as in Theorem 4 of (Chen et al., 2012). The resultant functional $F_1 : \mathfrak{lsc}_{\mathfrak{m}} \to \mathfrak{lsc}_{\mathfrak{m}}$ defined by $F_1(\rho)(s,s') = \max_{a\in A} \mathcal{K}(\rho)(P_a(s), P_a(s'))$ has a least fixed-point ρ_1^* given by the Knaster-Tarski fixed-point theorem. In fact, in this case the least fixed-point is the everywhere zero pseudometric - unsurprising since in our current setup all actions are allowable in all states and the reward parameter is the only feature that distinguishes states. But how might we interpret such a result more generally? In fact, there is some relation to the infinite-horizon average reward optimality criterion.

Let $\mathcal{M} = (S, \mathcal{B}_S, A, (P_a)_{a\in A}, r)$ be an MDP with the image of r contained in $[0,1]$ and recall the terminology of Section 2.2. The following definitions can be found in Chapter 5 of (Hernández-Lerma & Lasserre, 1996). Let $\pi \in \Pi$ be a policy on $\mathcal{M}$. Let $n \in \mathbb{N}$. The *n-stage value function for π* is defined by $J_n(\pi)(s) = \mathbb{E}_s^\pi[\sum_{t=0}^{n-1} r(a_t, x_t)]$ for all $s \in S$, the *average cost value function for π* by $J(\pi)(s) = \limsup_{n\to\infty} \frac{1}{n} J_n(\pi)(s)$ for all $s \in S$, and the average reward optimal value function by $J^*(s) = \sup_{\pi\in\Pi} J(\pi)(s)$ for all $s \in S$. The solution to the average reward Markov decision problem is a policy π^* such that $J(\pi^*) = J^*$.

Let us assume that $\mathcal{M}$ is finite, i.e., S is finite. The following can be found in Chapter 8 of (Feinberg & Shwartz, 2002), in particular Theorem 8.1, listed below as Theorem 5.1. A policy $\pi \in \Pi$ is said to be *Blackwell optimal* if and only if there exists $\gamma_0 \in (0,1)$ such that π is γ-optimal for all $\gamma \in (\gamma_0, 1)$.

Remark 2. *A stationary Blackwell optimality policy is also average reward optimal, and for such a policy π^*, $\lim_{\gamma\uparrow 1}(1-\gamma)V_\gamma(\pi^*) = J(\pi^*)$.*

Theorem 5.1. *In a finite MDP there exists a stationary Blackwell optimal policy.*

Let $K \in \Lambda(P,P)$ and $V_c^*(K)$ be the optimal value function for the MDP $\mathcal{M}(K)$ defined in Section 3 with the reward parameter θ^c scaled by $(1-c)^{-1}$. Then there exists a $K_c^* \in \Lambda(P,P)$, depending on c, such that $\rho_c^* = (1-c)V_c^*(K_c^*)$. It follows that $\lim_{c\uparrow 1}\rho_c^* = \lim_{c\uparrow 1}(1-c)V_c^*(K_c^*) \leq \lim_{c\uparrow 1}(1-c)V_c^*(K) \leq J^*(K)$ for any $K \in \Lambda(P,P)$. Thus, $\lim_{c\uparrow 1}\rho_c^* \leq \inf_{K\in\Lambda(P,P)} J^*(K)$. It remains to be seen whether or not the inequality is strict.

In the general case, the situation is much more complicated. For example, under a variety of conditions not listed here, Lemma 10.4.3 of (Hernández-Lerma & Lasserre, 1999) states the following.

Lemma 5.2. *There exists a constant α such that $\alpha = \limsup_{\gamma\uparrow 1}(1-\gamma)V_\gamma^* \leq J^*$.*

It follows that under the same conditions $\limsup_{c\uparrow 1}\rho_c^* \leq \alpha \leq \inf_{K\in\Lambda(P,P)} J^*(K)$. If equality were to hold in this case, then we would have for some $x \in S$, $\alpha = \limsup_{c\uparrow 1}\rho_c^*(x,x) = 0$, so that $\lim_{c\uparrow 1}\rho_c^*$ exists and is everywhere zero, i.e. the resulting equivalence would identify all states. Aside from the unspecified conditions, the distinction with the finite case is that $\lim_{c\uparrow 1}\rho_c^*$ need not exist to begin with.

Similarly, consider the expected total reward criterion. Here we might take the set of lower semicontinuous pseudometrics on S as in Theorem 2.5, but this time bounded with respect to a weighted supremum norm $\|\cdot\|_w$ for some weight function $w : S \times S \to [1,\infty)$. Thus, an unbounded function f that has a bounded norm with respect to w has its rate of growth bounded by w. Define the functional F by

$$F(\rho)(s,s') = \max_{a\in A}[\theta_a^0(s,s') + \mathcal{K}(\rho)(P_a(s),P_a(s'))].$$

Then if conditions are imposed so that the set of w-bounded lower semicontinuous pseudometrics on S is closed under F, it will have a least fixed-point (extended) pseudometric again corresponding to some expected total reward optimal value function. This line of research is very preliminary.

The major concern of this work along with (Ferns et al., 2014) is to clarify and unify results about the theory of bisimulation metrics in order to provide new avenues of attack for practical applications. An ongoing research goal is to find a more easily computable equivalence than that given by the current bisimulation metrics while maintaining as much as possible the theoretical guarantees. As far as estimating the given family of bisimulation metrics, the current interpretation as optimal value functions suggests that the most promising approaches involve Monte Carlo techniques, as in (Ferns et al., 2006), (Ferns et al., 2011), and most recently in (Comanici et al., 2012), or advanced approximate linear programming techniques as in (Pazis & Parr, 2013). More to the point, our strong intuition is that state-of-the-art methods for efficient reinforcement learning can be leveraged to develop state-of-the-art methods for efficient bisimulation metric computation, and vice versa. Very interesting recent work in this direction have been done by Bacci et al. (2013), who use greedy heuristics and decomposition techniques to speed up the computation of bisimulation metrics for MDPs. Computational approaches of this flavour should be further investigated.

In order for this approach to be really useful in practice, however, a few topics need to be further addressed by future work.

First, this work is highly dependent on couplings and the coupling method, though we have only just touched upon the subject. The study of couplings in theory and in practice is vast, and a proper discussion is beyond the scope of this work. A good source for the theory of couplings is (Lindvall, 2002). Moreover, as noted in (Chen et al., 2012), it is already known in the discrete case that the set of couplings (called *matchings* in that work) for two probability functions forms a polytope; and that optimizing a linear function over it amounts to optimizing over the finitely many vertices of the polytope (as is done in computing the discrete Kantorovich metric). We hope that this structure can be exploited to improve our theoretical result.

Additionally, we mentioned that the initial applications of bisimulation to MDPs exploited factored or structured representations. It would be fruitful to explore whether or not bisimulation metric reasoning principles can be applied to factored representations without having to flatten the state space. More generally, applying bisimulation metrics to the problem of constructing function approximators for MDP value functions is a very promising future direction, recent work (Comanici & Precup, 2012) has leveraged such metrics to tackle the problem of automatically generating features for function approximation.

Lastly, let us consider the problem of knowledge transfer between MDPs. Suppose $\mathcal{M}_X = (X, \mathcal{B}_X, A, P_X, r(X))$ and $\mathcal{M}_Y = (Y, \mathcal{B}_Y, A, P_Y, r(Y))$ are two MDPs with rewards in the unit interval and that $c \in (0,1)$ is a discount factor. For $K \in \Lambda(P_X, P_Y)$, consider the coupled MDP $\mathcal{M} = (X \times Y, \mathcal{B}_{X\times Y}, A, K, \theta)$ where $\theta_a(x,y) = (1-c)|r_a(X)(x) - r_a(Y)(y)|$ for all $x \in X$ and $y \in Y$. Does $\inf_{K\in\Lambda(P_X,P_Y)} V_c^*(K)(x,y)$ measure the bisimilarity of states x and y? Clearly, there is much work to be done to answer this question.

Acknowledgements

This work is dedicated with love to Prakash Panangaden in honour of his 60th birthday.

References

Abate, A. (2012). Approximation Metrics based on Probabilistic Bisimulations for General State-Space Markov Processes: a Survey. *Electronic Notes in Theoretical Computer Sciences*.

Bacci, G., Bacci, G., Larsen, K. G., & Mardare, R. (2013). Computing behavioral distances, compositionally. *Proceedings of the 38th International Symposium on Mathematical Foundations of Computer Science (MFCS)* (pp. 74–85).

Boutilier, C., Dearden, R., & Goldszmidt, M. (2000). Stochastic Dynamic Programming with Factored Representations. *Artificial Intelligence*, *121*, 49–107.

Chen, D., van Breugel, F., & Worrell, J. (2012). On the Complexity of Computing Probabilistic Bisimilarity. *FoSSaCS* (pp. 437–451). Springer.

Comanici, G., Panangaden, P., & Precup, D. (2012). On-the-Fly Algorithms for Bisimulation Metrics. *QEST* (pp. 94–103). IEEE Computer Society.

Comanici, G., & Precup, D. (2012). Basis Function Discovery Using Spectral Clustering and Bisimulation Metrics. *AAAI*.

Dean, T., Givan, R., & Leach, S. (1997). Model Reduction Techniques for Computing Approximately Optimal Solutions for Markov Decision Processes. *UAI* (pp. 124–131).

Desharnais, J., Jagadeesan, R., Gupta, V., & Panangaden, P. (2002). The Metric Analogue of Weak Bisimulation for Probabilistic Processes. *LICS* (pp. 413–422). IEEE Computer Society.

Desharnais, J., Laviolette, F., & Zhioua, S. (2013). Testing Probabilistic Equivalence Through Reinforcement Learning. *Information and Computation*, *227*, 21–57.

Doberkat, E.-E. (2007). *Stochastic Relations. Foundations for Markov Transition Systems.* Chapman & Hall/CRC.

Feinberg, E., & Shwartz, A. (Eds.). (2002). *Handbook of Markov Decision Processes - Methods and Applications*. Kluwer International Series.

Ferns, N., Castro, P. S., Precup, D., & Panangaden, P. (2006). Methods for Computing State Similarity in Markov Decision Processes. *UAI*.

Ferns, N., Panangaden, P., & Precup, D. (2004). Metrics for Finite Markov Decision Processes. *UAI* (pp. 162–169).

Ferns, N., Panangaden, P., & Precup, D. (2005). Metrics for Markov Decision Processes with Infinite State Spaces. *UAI* (pp. 201–208).

Ferns, N., Panangaden, P., & Precup, D. (2011). Bisimulation Metrics for Continuous Markov Decision Processes. *SIAM Journal on Computing*, *40*, 1662–1714.

Ferns, N., Precup, D., & Knight, S. (2014). Bisimulation for Markov Decision Processes Through Families of Functional Expressions. *Horizons of the Mind. A Tribute to Prakash Panangaden* (pp. 319–342). Springer.

Folland, G. B. (1999). *Real analysis: Modern techniques and their applications*. Wiley-Interscience. Second edition.

Giry, M. (1982). A Categorical Approach to Probability Theory. *Categorical Aspects of Topology and Analysis*, 68–85.

Givan, R., Dean, T., & Greig, M. (2003). Equivalence Notions and Model Minimization in Markov Decision Processes. *Artificial Intelligence*, *147*, 163–223.

Hernández-Lerma, O., & Lasserre, J. B. (1996). *Discrete-Time Markov Control Processes : Basic Optimality Criteria*. Applications of Mathematics. Springer.

Hernández-Lerma, O., & Lasserre, J. B. (1999). *Further Topics on Discrete-Time Markov Control Processes*. Applications of Mathematics. Springer.

Larsen, K. G., & Skou, A. (1991). Bisimulation Through Probabilistic Testing. *Information and Computation*, *94*, 1–28.

Li, L., Walsh, T. J., & Littman, M. L. (2006). Towards a Unified Theory of State Abstraction for MDPs. *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics* (pp. 531–539).

Lindvall, T. (2002). *Lectures on the Coupling Method*. Dover Publications Inc.

Pazis, J., & Parr, R. (2013). Sample Complexity and Performance Bounds for Non-Parametric Approximate Linear Programming. *AAAI*.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc.

Srivastava, S. M. (2008). *A Course on Borel Sets*, vol. 180 of *Graduate texts in mathematics*. Springer.

Sutton, R. S., & Barto, A. G. (2012). *Reinforcement Learning: An Introduction (Second Edition, In Progress)*. MIT Press.

van Breugel, F., & Worrell, J. (2001a). Towards Quantitative Verification of Probabilistic Transition Systems. *ICALP* (pp. 421–432). Springer.

Villani, C. (2003). *Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58)*. American Mathematical Society.