State-Similarity Metrics for Continuous Markov Decision Processes

Norman Francis Ferns

Doctor of Philosophy

Reasoning and Learning Lab
School of Computer Science

McGill University

Montreal,Quebec

October 2007

A thesis submitted to McGill University
in partial fulfilment of the requirements of
the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-50813-8
Our file Notre référence
ISBN: 978-0-494-50813-8

DEDICATION

For

Lauren, *Ryan*, *Raquel*, *Ruth*, and *Norman*

ACKNOWLEDGEMENTS

The list of those to whom I am in debt is a long one; here are but a few ...

Firstly, I would like to thank the bestest supervisors in the whole wide world, Prakash Panangaden and Doina Precup. Smarter people I do not know. You have been nothing short of kind and supportive, more patient and understanding with me than warranted, and just all-around swell. I really do appreciate everything. Thank you.

Additionally, I would like to thank Joelle Pineau and Dmitry Jakobson; for their open-door policies, for always lending a helping hand, whether on the topic of POMDPs or semicontinuity, or coming to the rescue with whatever insane request I brought to them. Thank you.

I must also thank the people in the School of Computer Science, in particular Diti and the rest of the lovely ladies in the CS office, for a wonderful learning atmosphere and cutting through the administrative red tape with ease and a smile.

This work would not be here if not for the help of two geniuses: my brother, Ryan Ferns, and super-scientist (and possibly Batman?), Marc Berndl. Thank you both for the many, many, many hours of helpful discussions!

A very special shout out goes to my best mates: my bro's Ian and Buds, Z.O.M.B.I.E., and my dearest Iepuraş. I love you guys. Thanks for being my friends ...

Lastly, my family - the people who would kill for me (or at the very least help me hide the body). I can never repay you for all you have done for me: the

emotional and financial support, the love, and more. Mum, you have sacrificed everything for me, and I will never forget that. Thank you. Dad, you have worked so hard to provide us with a home for all these years. Thank you. Raquel, you are an inspiration. Really. I have always tried to do at least as well as you. I admire the way you continue to grow as a person. You are beautiful. Thank you for looking out for me. Ryan, you are the most sweetest wonderful loving caring person I have ever known. You have and always will be like a son to me. Thank you for being in my life. And to Lauren, my partner in crime: only the smartest, most beautiful girl in the world could have gotten me to finish this thesis! Thank you for all your love and support. You are amazing and I love you ♡

ABSTRACT

In recent years, various metrics have been developed for measuring the similarity of states in probabilistic transition systems (Desharnais et al., 1999; van Breugel & Worrell, 2001a). In the context of Markov decision processes, we have devised metrics providing a robust quantitative analogue of bisimulation. Most importantly, the metric distances can be used to bound the differences in the optimal value function that is integral to reinforcement learning (Ferns et al. 2004; 2005). More recently, we have discovered an efficient algorithm to calculate distances in the case of finite systems (Ferns et al., 2006). In this thesis, we seek to properly extend state-similarity metrics to Markov decision processes with continuous state spaces both in theory and in practice. In particular, we provide the first distance-estimation scheme for metrics based on bisimulation for continuous probabilistic transition systems. Our work, based on statistical sampling and infinite dimensional linear programming, is a crucial first step in real-world planning; many practical problems are continuous in nature, e.g. robot navigation, and often a parametric model or crude finite approximation does not suffice. State-similarity metrics allow us to reason about the quality of replacing one model with another. In practice, they can be used directly to aggregate states.

SOMMAIRE

Au cours des dernières années, plusieurs métriques ont été développés pour mesurer l'équivalence des états dans les systèmes de transition probabilistes (Desharnais et al., 1999; van Breugel & Worrell, 2001a). Pour les processus de décision Markoviens, nous avons créé des métriques qui fournissent un analogue quantitatif de la bisimulation et tels que les diffèrences dans la fonction valeur optimale de « l'apprentissage par renforcement » sont moindres que les distances métriques (Ferns et al. 2004; 2005). Plus récemment, nous avons découvert un algorithme rapide pour calculer les distances dans les systèmes finis (Ferns et al., 2006). Ici, nous espérons développer ces travaux pour des processus de décision Markoviens dans lesquels l'espace des états est continu. En particulier, nous fournissons le premier algorithme pour calculer les métriques pour les systèmes de transition probabilistes continus, en utilisant des techniques statistiques et de programmation linéaire en dimensions infinies. Notre travail est une première étape cruciale dans l'apprentissage par renforcement pour des problèmes réalistes: beaucoup de problèmes sont naturellement continus, par exemple, la navigation d'un robot mobile, et souvent un modèle paramétrique ou une approximation finie imprécise ne suffit pas. Nos distances nous permettent d'évaluer la qualité de remplacer un modèle avec d'autres. De plus, elles peuvent être employées directement pour l'agrégation des états.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Motivations

Markov decision processes (MDPs) are the standard mathematical model of choice when it comes to sequential decision making under uncertainty (Boutilier et al., 1999). The objective of this decision making is to maximize a cumulative measure of long-term performance, called the *return*. Standard dynamic programming algorithms such as value iteration or policy iteration (Puterman, 1994) allow one to compute the optimal expected return for any state, and in turn, the optimal method of decision making, the optimal policy, that generates this return. However, in many practical situations the state space of an MDP may be too large, possibly even infinite, in which case the standard algorithms cannot be applied. Similarly, MDPs with a high degree of stochasticity, that is, when there are many possible outcome states for probabilistic state transitions, can be much more problematic to solve than those that are nearly deterministic (Likhachev et al., 2005). Thus, one usually turns to approximation theory to find a simpler relevant model; the hope is that this can be done in such a manner as to construct an "essentially equivalent" MDP with drastically reduced complexity, thereby allowing the use of classical solution methods while at the same time providing a guarantee that solutions to the reduced MDP can be extended to the original model.

Recent MDP research on defining equivalence relations on MDPs (Dean and Givan 1997; 2003) has built on the notion of strong probabilistic bisimulation from concurrency theory. Probabilistic bisimulation was introduced by Larsen and Skou (1991) based on bisimulation for nondeterministic systems due to Park (1981) and Milner (1980). In a probabilistic setting, bisimulation can be described as an equivalence relation that relates two states precisely when they have the same probability of transitioning to classes of equivalent states. The extension of bisimulation to transition systems with rewards was carried out in the context of MDPs by Givan et al. (2003) and in the context of performance evaluation by Bernardo and Bravetti (2003). In both cases, the motivation is to use the equivalence relation to aggregate the states and get smaller state spaces. The basic notion of bisimulation is modified only slightly by the introduction of rewards.

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

A natural remedy is to use (pseudo)metrics, as metrics are natural quantitative analogues of equivalence relations. The triangle inequality, for example, can be interpreted as a quantitative generalization of the axiom of transitivity: if states x and y, and y and z, are close in distance then so too must be states x and z.

The metrics on which we focus in this work specify the degree of similarity of a system's states, with a distance of zero corresponding to exact equivalence, or bisimulation. Based on work in the context of labeled Markov processes (Desharnais et al. 1999; 2004; van Breugel and Worrell 2001a; 2001b), we sought to extend bisimulation for MDPs quantitatively in terms of such metrics (Ferns et al. 2004; 2005; 2006). In the case of infinite state spaces, we were able to prove existence of state-similarity metrics satisfying certain continuity conditions. For finite models the situation is even better; we discovered an efficient method for estimating the distances using techniques from statistical sampling and network optimization. Still, the full potential of the state-similarity metrics has yet to be realized.

Consider now the following scenario: a robot is set to navigate a foreign terrain. After brief exploration, data is collected and a probabilistic planning model is devised. The problem is most naturally modeled as a continuous state space Markov decision process. One of the following occurs:

- Several finite state approximations are proposed; which of these behaves most like the original?
- Several deterministic approximations are proposed; which of these behaves most like the original?
- A package of planning strategies has been precomputed for each of several known models in similar planning tasks. We would like to reuse these strategies. From which model should the robot choose its planning strategies?

- The decision maker decides to aggregate states and plan with the resulting finite state model; how should states be aggregated so as to minimize planning error while maintaining model accuracy?

Alternatively, consider that the problem is most naturally modeled by a partially observable Markov decision process. It is well-known that this can be equivalently modeled as a continuous state (belief) space Markov decision process. Again, how do we proceed?

These are the questions we hope to answer through this work. Robot navigation is just one of many real-world problems that are most naturally modeled with a continuous state space. State-similarity metrics in theory allow us to reason about proposed approximation schemes. More importantly, in practice we can use them to aggregate states, to assess the quality of an aggregation, to assess the quality of several finite state approximation schemes, to bound the error in using solutions from similar models, and in particular, in using deterministic models. Such work is crucial, especially in those cases where a parametric model or crude finite approximation will not suffice.

1.2 Contributions

Here we unify and strengthen the results of Ferns et al. (2005) for infinite state MDPs by providing state-similarity metrics with better continuity properties. The most important contribution of this thesis, however, is an extension of the sampling algorithm used in the finite case (Ferns et al., 2006) to MDPs whose state spaces are compact metric spaces; in short, one can effectively estimate state-similarity distances everywhere by estimating their values on a finite set. Crucial to this distance-approximation scheme is a novel application of a uniform Glivenko-Cantelli theorem, essentially guaranteeing uniform convergence of empirical probability measures to the true measures.

Specifically, the main original contributions of this thesis are the following:

- We extend an approach to bisimulation metrics for finite state probabilistic transition systems due to van Breugel and Worrell (2001b), based on linear programming, to bisimulation metrics for continuous state space Markov decision processes using infinite dimensional linear programming (theorem 3.4.2). This is a refinement of previous work (Ferns et al., 2005).
- We prove Lipschitz continuity of the optimal value function with respect to our bisimulation metrics for continuous state space Markov decision processes (theorem 3.4.10). This is a refinement of previous work (Ferns et al., 2005).
- Our key original result is the stochastic distance-approximation scheme based on the assignment problem from linear programming (theorem 4.1.4). The entirety of Chapter 4 is original work.

1.3 An Outline

The thesis is organized as follows: in the next chapter, we present a brief mathematical survey and then review finite Markov decision processes, including a discussion of the standard reinforcement learning paradigm, bisimulation, bisimulation metrics and methods for computing these. We also take a look at related work. Chapter 3 then shifts the discussion to infinite state spaces, introducing issues of measurability and continuous analogues of concepts introduced in the previous chapter. We delve into the details of the Kantorovich functional, an infinite linear program that can be used to define a metric on probability measures. This, in turn, is used in the first major result: existence of state-similarity metrics generalizing bisimulation, along with several continuity properties. We conclude with an important reinforcement learning bound and a simple calculation, illustrating the use of metric reasoning. In Chapter 4 we present our central result: an approximation scheme for estimating distances for MDPs whose state spaces are compact metric spaces. We review the uniform Glivenko-Cantelli property and apply it in conjunction with a fixed point theorem to arrive at our Monte-Carlo algorithm. Several error bounds are provided and in Chapter 5 we provide some illustrations of the algorithm in practice. In the concluding chapter, we present a summary of our results and discuss directions for future research.

CHAPTER 2
Background

In this chapter we will review some fundamental mathematical background as well as the basics of Markov decision processes with respect to reinforcement learning, bisimulation, and quantitative state-similarity along with current methods for its computation. For the sake of convenience, this latter material will initially be presented in the context of finite systems; the issues that arise in moving to infinite state spaces will be discussed in the following chapters. We conclude by looking at related work in the field.

2.1 A Mathematical Review

We begin with a brief mathematical survey of the results that are most relevant to this thesis, and in particular, the mathematics of continuous systems. Results will be stated without proof and can be found in most classical texts in probability and analysis, such as Rudin (1976), Folland (1999), Dudley (2002), and Billingsley (1968).

2.1.1 Metric Spaces

A metric is perhaps the simplest geometric structure that one can impose on a space. It is essentially a distance function, that is, a means of assigning a nonnegative numerical weight to pairs of points on a set in order to quantify how far apart they are. More formally we say a *pseudometric* on S is a map $\rho : S \times S \to [0, \infty)$ such that for all s, s', s'' in S:

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

If the converse of the first axiom holds as well, we say ρ is a *metric.*[1]

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

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

2.1.1.1 Convergence

A metric easily allows one to speak of the convergence of elements in a space: a sequence converges to a limit point if the distance between that limit point and the points in the sequence can eventually be made arbitrarily small. Formally, a sequence of elements $\{x_n\}$ converges to an element x in a metric space (S, ρ) if and only if for every positive ϵ there exists a natural number N depending on ϵ such that for all $n \geq N$, $\rho(x_n, x) < \epsilon$.

As an example, whenever we speak of a sequence of real-valued functions *converging uniformly*, we are implicitly invoking convergence in the space of

[1] For convenience, we will use the terms metric and pseudometric interchangeably, though we really mean the latter.

bounded real-valued functions equipped with the metric induced by the uniform norm, i.e. $\|f\| := \sup_{x \in S} |f(x)|$.

Sometimes it is convenient to speak of the convergence of a sequence without having a definite candidate for its limit in mind. Suppose instead that we had considered a sequence whose pairwise distances could eventually be made arbitrarily small; we might expect that the sequence itself should converge. Unfortunately, such is not always the case. Formally, a sequence $\{x_n\}$ is said to be *Cauchy* if for every positive ϵ there exists a natural number N depending on ϵ such that for all $n, m \geq N$, $\rho(x_n, x_m) < \epsilon$. A metric space in which every Cauchy sequence converges is said to be *Cauchy-complete* or simply *complete.* For example, let $\mathfrak{met}$ be the set of bounded pseudometrics on S equipped with the metric induced by the uniform norm, $\|h\| = \sup_{s,s'} |h(s, s')|$. Then $\mathfrak{met}$ is a complete metric space.

2.1.1.2 Special Sets

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

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

Let us note that different metrics can produce the same collection of open sets on a space, and that some properties depend only on this collection of open sets, rather than on a given metric. The set S equipped with a given collection of open sets is called a *topological space.* Specifically, a collection $\mathcal{T}$ of subsets of S forms a *topology* on S if and only if:

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

Properties that refer only to the collection of open sets will be referred to as topological.

2.1.1.3 Continuity

Continuity is a crucial property for our work on approximating spaces and functions on those spaces. Loosely speaking, a function is continuous if the output of the function cannot change too abruptly with small changes in its input. More formally, a function $f : (X, \rho_X) \to (Y, \rho_Y)$ is *continuous at a point* $x \in X$ if for every $\epsilon > 0$ there is a $\delta > 0$, depending on x and ϵ, such that for all $x' \in X$ with $\rho_X(x, x') < \delta$ we have $\rho_Y(f(x), f(x')) < \epsilon$. We say f is *continuous* if it is continuous at every point of X. If δ can be chosen so as to depend on ϵ alone, i.e. independent of the point x, then f is said to be *uniformly continuous.* A very strong form of uniform continuity is Lipschitz continuity: f is *Lipschitz continuous* if for some constant α, $\rho_Y(f(x), f(x')) \leq \alpha\rho_X(x, x')$ for all $x, x' \in X$. The constant α is known as the *Lipschitz constant* for this mapping, though some take this term to refer to the greatest lower bound of all possible constants. We will sometimes write that f is α-Lipschitz continuous. Obviously every Lipschitz continuous function is uniformly continuous, and every uniformly continuous function is continuous, but the converse is not generally true in either case. For compact metric spaces, however, the situation is much more well-behaved. Here, every continuous function is indeed uniformly continuous. Moreover, if f is real-valued then it has a minimum value and a maximum value, each of which is attained.

Continuity in metric spaces can alternatively be characterized in terms of convergent sequences: f is continuous if for every convergent sequence $\{x_n\}$ in X with limit x, the sequence $\{f(x_n)\}$ is convergent with limit $f(x)$. One can use this to loosen the definition of continuity in several ways. One generalization that is useful in this work is semicontinuity. Formally, a real-valued function f on a metric space (X, ρ) is *lower semicontinuous* if for any sequence $\{x_n\}$ converging to x in X, $\liminf_{n\to\infty} f(x_n) \geq f(x)$; one can analogously define f to be *upper semicontinuous* by requiring $\limsup_{n\to\infty} f(x_n) \leq f(x)$. It is easily seen that a real-valued function is continuous if and only if it is both lower semicontinuous and upper semicontinuous. The intuition behind these definitions is that semicontinuous functions allow for abrupt (discontinuous) jumps in one vertical direction; this can be seen through the prototypical examples of semicontinuous functions: the indicator function of an open set is always lower semicontinuous while the indicator function of a closed set is always upper semicontinuous. In this work, we will be particularly interested in lower semicontinuous functions due to several important properties; for example, the pointwise supremum of an arbitrary collection of uniformly bounded lower semicontinuous functions on a Polish metric space is itself lower semicontinuous, and a lower semicontinuous function on a compact subset of a Polish metric space attains its minimum at some point in the subset.

Continuity in topological spaces is defined as follows: a function $f : (X, \mathcal{T}_X) \rightarrow (Y, \mathcal{T}_Y)$ is continuous if for each open set $O_Y \in \mathcal{T}_Y$, the preimage $f^{-1}(O_Y) \in \mathcal{T}_X$. If the topologies $\mathcal{T}_X$ and $\mathcal{T}_Y$ are induced by metrics, then this definition actually coincides with the metric definition of continuity. Continuity is important for defining equivalence of topological spaces; two topological spaces are equivalent, or *homeomorphic*, if there exists a continuous bijection between them such that its inverse is also continuous. A *Polish space*, for example, is any topological space that is homeomorphic to a Polish metric space, as defined above. Another important topological space is an *analytic space.* An analytic space is the continuous image of a Polish space under a map between Polish spaces.

2.1.2 Fixed Points

Fixed point theory plays a major role in this thesis. Here we recall some basic definitions and a theorem from fixed point theory on lattices, which can be found in any basic text (Winskel, 1993).

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

Theorem 2.1.1 (Knaster-Tarski Fixed Point Theorem). *Let L be a complete lattice, and suppose $f : L \rightarrow L$ is monotone. Then f has a least fixed point, which* *is also its least prefixed point, and f has a greatest fixed point, which is also its greatest postfixed point.*[2]

This theorem will later be applied to two rather interesting complete lattices $\mathfrak{Rel}$ and $\mathfrak{Equ}$, the space of binary relations and the space of topologically-closed equivalence relations on a Polish space S, respectively. Here, we equip each with the subset ordering, clearly obtaining partial orders. The greatest lower bound of a set of relations is simply their intersection. The same can be said for a set of equivalence relations - and moreover, an arbitrary intersection of topologically closed sets is topologically closed. Hence, both spaces are complete lattices.[3]

A more common fixed point theorem comes from the theory of metric spaces and has the advantage of being constructive in nature; its proof can be found in most basic texts in analysis, e.g. (Rudin, 1976).

Theorem 2.1.2 (Banach Fixed Point Theorem). *Suppose (X, d) is a complete metric space and $T : X \to X$ is a contraction mapping, i.e. for some $c \in [0, 1)$*

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

for all x, x' in X. Then:

1. *T has a unique fixed point, x^*, and*

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

[3] Existence of least upper bounds follows from that of greatest lower bounds.

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

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

2.1.3 Probability and Measure

A rather unfortunate consequence of moving to uncountably infinite state spaces is that we can no longer specify transition probabilities point-to-point; one needs to specify probabilities on sets of points and even then not all sets can be "measured" in this way. Formally, we say a *σ-algebra* or *σ-field* on S is a collection Σ of subsets of S satisfying the following axioms:

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

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

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

1. $\mu(\emptyset) = 0$, and
2. for any pairwise disjoint collection of sets $\{E_i\}$ in Σ, $\mu(\bigcup E_i) = \Sigma \mu(E_i)$.

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

Sometimes we need to assign weights of a probabilistic type to all subsets of a space, at the cost of satisfying all the nice properties of a probability measure; such is frequently the case in the theory of empirical processes, where one cannot guarantee that all the sets one may encounter in practice will be measurable. An *outer probability measure* is a set function $\phi : 2^S \longrightarrow [0,1]$ satisfying

1. $\phi(\emptyset) = 0$,
2. $E \subset F$ implies $\phi(E) \leq \phi(F)$, and
3. for any countable collection $\{E_i\}$ of subsets of S, $\phi(\bigcup E_i) \leq \Sigma \phi(E_i)$.

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

A probability measure on a metric space is *tight*, or *inner regular*, if it can be approximated from within by compact sets, i.e. μ is tight if for every Borel measurable set E, $\mu(E) = \sup_K \mu(K)$ where the supremum is taken over all compact subsets K contained in E. Every probability measure on a Polish metric space is tight; this is known as *Ulam's Tightness Theorem.*

Measures can be extended to act on functions through the process of integration. We will assume the reader is familiar with the basic ideas of integration, if not the details, as the details are involved and add nothing to the exposition here. Suffice it to say that, just as only certain subsets can be measured, so too can only certain functions be integrated. Formally, a function f between measurable spaces (X, Σ_X) and (Y, Σ_Y) is said to be *measurable* if the preimage of every

Σ_Y-measurable set is Σ_X-measurable, i.e. $\{f^{-1}(E) : E \in \Sigma_Y\} \subseteq \Sigma_X$. A real-valued function f on a measurable space (S, Σ) is measurable, or in the language of probability theory, a *random variable*, if it is measurable as just defined, where $\mathbb{R}$ is equipped with its usual Borel σ-field. The prototypical measurable functions are the *simple functions*: finite linear combinations of indicator functions on measurable sets. If S is a metric space and Σ its Borel σ-field, then every continuous function on S is measurable. Given a sequence of measurable functions, its pointwise supremum, infimum, and limit (when it exists) are all measurable. Lastly, if the integral of the absolute value of a measurable function f with respect to a measure μ exists and is finite, then f is said to be *integrable*. The collection of all such f for a given μ is denoted by $L^1(\mu)$ (here it is standard to identify functions that differ on a set of μ-measure zero).

2.1.3.1 Weak Convergence of Probability Measures

Let us now consider convergence of probability measures on a metric space. Since probability measures are essentially just set functions, it is natural to attempt to analyze their convergence properties through pointwise converge, i.e. to say that a sequence of probability measures $\{\mu_n\}$ converges to probability measure μ if $\{\mu_n(E)\}$ converges to $\mu(E)$ for every measurable set E. However, such convergence is too strong: consider the Dirac measure δ_x, which assigns a value of 1 if and only if a given measurable set contains the point x and 0 otherwise. Take $[0, 1]$ with its Borel σ-algebra and consider the sequence of Dirac measures on $\{\frac{1}{n} : n \in \mathbb{N}\}$. It would be quite natural to expect, if not demand, that this sequence converges to the Dirac measure at zero. However, taking the Borel measurable singleton $\{0\}$ in the definition of pointwise convergence would yield $\lim_{n\to\infty} \delta_{\frac{1}{n}}(\{0\}) = 0 = \delta_0(\{0\}) = 1$, which is clearly not the case. It is not hard to show here that pointwise convergence over the measurable sets is equivalent to pointwise convergence over bounded measurable functions, i.e. convergence of $\{\mu_n(f)\}$ to $\mu(f)$ for every bounded measurable function f. Therefore, one way of weakening convergence is to consider a similar pointwise convergence, but over a smaller class of functions. Formally, we say that $\{\mu_n\}$ *converges weakly* to μ if $\{\mu_n(f)\}$ converges to $\mu(f)$ for every bounded *continuous* real-valued function f. It is not hard to show that the Dirac measures on $\{\frac{1}{n} : n \in \mathbb{N}\}$ do indeed converge weakly to the Dirac measure at 0.

2.1.3.2 Empirical Processes

Consider now an ambient probability space $(\Omega, \mathcal{A}, \mathbb{P})$ over which we sample n points $\{X_1, X_2, \ldots, X_n\}$ with values in (S, Σ) independently and with identical distribution μ, i.e. each X_i is a measurable map from $(\Omega, \mathcal{A}, \mathbb{P})$ to (S, Σ) such that $\mathbb{P}(\{\omega \in \Omega : X_i(\omega) \in E\}) = \mathbb{P}(X_i^{-1}(E)) = \mu(E)$ for every E in Σ. Define the nth *empirical probability measure* μ_n of μ to be the average of the Dirac measures at each X_i, i.e. $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Each μ_n is in effect a random measure; that is, for each $\omega \in \Omega$, $\mu_n(\omega) := \frac{1}{n}\sum_{i=1}^n \delta_{X_i(\omega)}$ is a probability measure. Does this sequence of random probability measures $\{\mu_n\}$ converge?

Recall that a sequence of random variables $\{Y_n\}$ converges to a random variable Y *in probability*, if for every $\epsilon > 0$,

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

and *almost surely* if

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

The Weak Law of Large Numbers (Strong Law of Large Numbers) tells us that for each real-valued bounded continuous f, the sequence of random variables $\{\mu_n(f)\} = \{\frac{1}{n}\sum_{i=1}^{n} f(X_i)\}$ converges to $\mu(f)$ in $\mathbb{P}$-probability ($\mathbb{P}$-almost surely). If the convergence was uniform over the set $\mathcal{F}$ of all bounded continuous functions, i.e. if $\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)|$ converged to zero , then it would follow that the empirical measures themselves converged weakly. This turns out to be a useful property in itself. Let us note that the maps $\omega \mapsto \sup_{f\in\mathcal{F}} |\mu_n(\omega)(f) - \mu(f)|$ need not be measurable since they involve taking a supremum over the possibly uncountable collection $\mathcal{F}$. Thus, we will need to use the outer probability $\mathbb{P}^*$ when speaking of their convergence. Any class $\mathcal{F}$ of integrable functions for μ such that $\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)|$ converges to zero in $\mathbb{P}^*$-outer probability ($\mathbb{P}^*$-almost surely) is said to be a *weak (strong) Glivenko-Cantelli class.* If $\mathcal{F}$ is a Glivenko-Cantelli class for every probability measure on (S, Σ) then it is said to be a *universal Glivenko-Cantelli class.* Lastly, if the rate of $\mathbb{P}^*$-convergence can be made to be uniform over all μ, i.e. if for every positive ϵ there exists a natural number N depending only on ϵ such that for all μ and all $n \geq N$, $\mathbb{P}^*(\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)| > \epsilon) < \epsilon$, then $\mathcal{F}$ is said to be a *uniform Glivenko-Cantelli class.*

We are now ready to proceed.

2.2 Reinforcement Learning

We will adhere to the somewhat simplified view that artificial intelligence is the science of intelligent agents, that is, the entities that perceive and act within an environment. The environment, then, is defined to be all that is external to the agent.

Accordingly, we define reinforcement learning (RL)[4] to be that branch of AI that deals with an agent learning through interaction with its environment in order to achieve a goal. The intuition behind reinforcement learning is that of learning by trial and error. By contrast, in supervised learning an external supervisor provides examples of desired behaviour from which an agent can learn, much as a student learns from a teacher.

Applications of reinforcement learning include optimal control in robotics (Lane & Pack Kaelbling, 2002), meal provisioning (Goto et al., 2004), scheduling, brain modelling, game playing, and more.

2.2.1 Markov Decision Processes

The interaction of an agent with its environment in reinforcement learning can be formally described by the Markov decision process framework below:

Consider the sequential decision model represented in Figure 2–1 (Sutton & Barto, 1998), depicting the interaction between a decision-maker, or agent, and its environment.

[4] This is also known as neuro-dynamic programming.

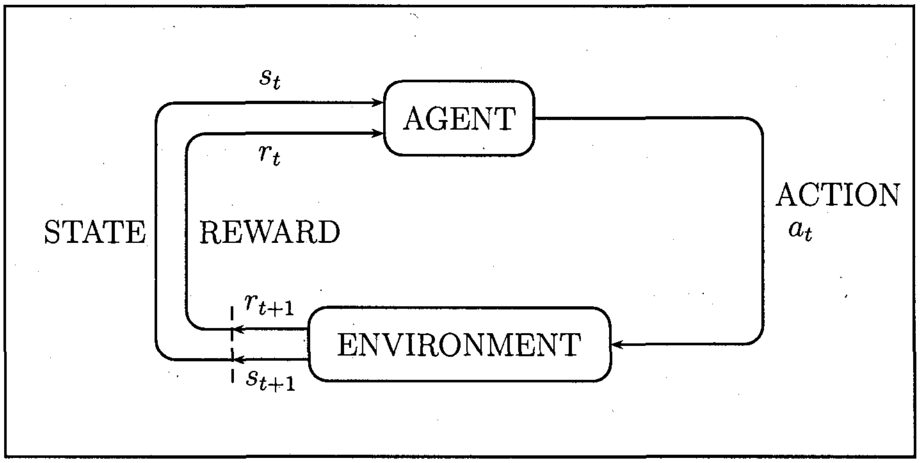


Figure 2–1: Agent-environment interaction

We assume that time is discrete, and that at each discrete time step $t \in \{0, 1, 2, \ldots, T\}$, the agent perceives the current state of the environment s_t from the set of all states S. We refer to T as the *horizon* and note that it may be either finite or infinite. On the basis of its state observation the agent selects an action a_t from the set of actions allowable in s_t, A_{s_t}. As a consequence, the following occurs immediately in the next time step: the agent receives a numerical signal r_{t+1} from the environment and the environment evolves to a new state s_{t+1} according to a probability distribution induced by s_t and a_t. The agent perceives state s_{t+1} and the interaction between agent and environment continues in this manner, either indefinitely or until some specified termination point has been reached, in accordance with the length of the horizon. Here, we think of r_{t+1} as a means of providing the agent with a reward or a punishment as a direct consequence of its own actions, thereby enabling it to learn which action-selection strategies are good and which are bad via its own behaviour.

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

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

and are *stationary*, i.e. independent of time:

$$\forall t, Pr(s_{t+1} = s'|s_t = s, a_t = a) = P^a_{ss'}$$

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

Definition 2.2.1. A *finite Markov decision process* is a quadruple

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

where:

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

We frequently take $A_s = A$, i.e. all actions are allowable in all states.

A finite Markov decision process (hereafter, MDP) can also be specified via a state-transition diagram; Figure 2–2, for example, depicts a finite MDP with 4 states and 1 action.

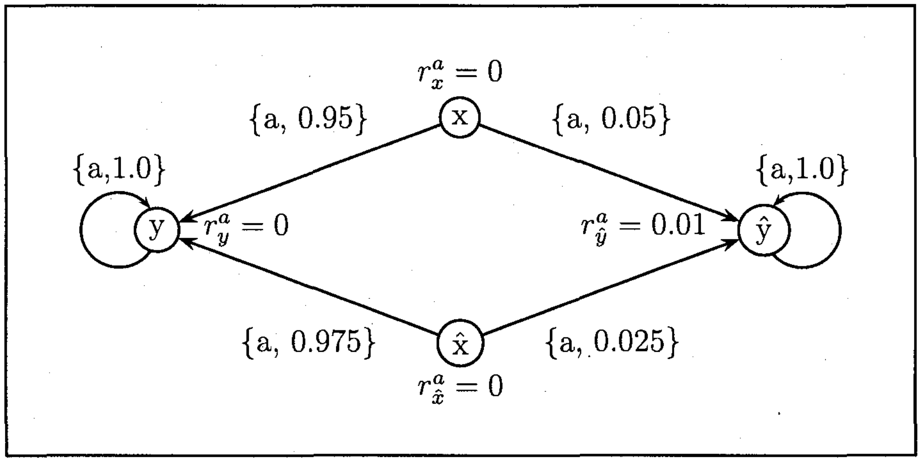


Figure 2–2: State transition diagram for a simple finite MDP

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

2.2.2 Policies and Optimality Criteria

An action selection strategy, or *policy*, is essentially a mapping from states to actions, i.e. a policy dictates what action should be chosen for each state. More generally, one allows for policies that are stochastic, history-dependent, and even non-stationary. Here we will restrict our attention to randomized stationary

Markov policies. Formally, a policy is a mapping $\pi : S \times A \longrightarrow [0, 1]$, such that $\pi(s, \cdot)$ is a probability distribution on A for each $s \in S$.

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

The total discounted reward criterion involves geometrically discounting the reward sequence. The intuition is that rewards obtained in the future are less valuable than rewards received immediately, an idea prevalent in economic theory. Alternatively, we may view it simply as a mathematical tool to ensure convergence. In any case, the discounted reward model possesses many nice properties, such as a simplified mathematics in comparison to other proposed optimality criteria and existence of stationary optimal policies (Puterman, 1994). For this reason, it is the dominant criterion used for RL tasks, and we concentrate on it in this work.

2.2.3 The Value of a Policy

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

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

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

$$\begin{aligned}
V^{\pi}(s) &= \mathbb{E}^{\pi}[R_t|s_t = s] = \mathbb{E}^{\pi}[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s] \\
&= \sum_{a\in A_s}\pi(s,a)\mathbb{E}^{\pi}[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s, a_t = a] \\
&= \sum_{a\in A_s}\pi(s,a)(r_s^a + \gamma\mathbb{E}^{\pi}[\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t = s, a_t = a]) \\
&= \sum_{a\in A_s}\pi(s,a)(r_s^a + \gamma\sum_{s'\in S}P_{ss'}^a\mathbb{E}^{\pi}[\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t = s, a_t = a, s_{t+1} = s']) \\
&= \sum_{a\in A_s}\pi(s,a)(r_s^a + \gamma\sum_{s'\in S}P_{ss'}^a\mathbb{E}^{\pi}[R_{t+1}|s_{t+1} = s']) \\
&= \sum_{a\in A_s}\pi(s,a)(r_s^a + \gamma\sum_{s'\in S}P_{ss'}^a V^{\pi}(s'))
\end{aligned}$$

The linear equations

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

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

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

$$V^*(s) = \max_{a \in A_s} (r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V^*(s')), \forall s \in S \tag{2.2}$$

of which it is the unique solution (see sections 6.1 and 6.2 of Puterman (1994)). These are known as the *Bellman optimality equations.*

It is worth remarking that the existence and uniqueness of the solutions in these Bellman equations can be obtained from the Banach Fixed Point Theorem, Theorem 2.1.2, by applying the appropriate contraction mapping over the space of bounded real-valued functions on S equipped with the metric induced by the uniform norm.

2.2.4 Value Iteration and Policy Iteration

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

Theorem 2.2.2 (Policy Evaluation). *Given a randomized stationary policy* π, *define*

- $V_0^\pi(s) = 0$, $\forall s \in S$ *and*

- $V_{i+1}^{\pi}(s) = \sum_{a \in A_s} \pi(s,a)(r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_i^{\pi}(s')), \forall i \in \mathbb{N}, \forall s \in S$.

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

Theorem 2.2.3 (Value Iteration). *Define*

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

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

These results allow one to compute value functions up to any prescribed degree of accuracy. For example, if one is given a positive ϵ then iterating until the maximum difference between consecutive iterates is $\frac{\epsilon(1-\gamma)}{2\gamma}$ guarantees that the current iterate differs from the true value function by at most ϵ (Puterman, 1994).

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

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

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

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

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

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

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

2.3 Discrete State Similarity

Let $(S, A, \{P^a_{ss'}\}, \{r^a_s\})$ be a given finite MDP. When should two states be placed in the same cluster of a state aggregation? Equivalently, what is the best state equivalence for MDP model reduction?

Givan et al. (2003) investigated several notions of MDP state equivalence for MDP model minimization: action-sequence equivalence, optimal value equivalence, and bisimulation. Two states are deemed action-sequence equivalent if for any fixed finite sequence of actions, their distributions over reward sequences are the same. Here let us remark that for any state, a fixed finite sequence of actions of length n induces a probability distribution over reward sequences of size n by means of the MDP's system dynamics. As Givan et al. note, the problem with action-sequence equivalence is that it may equate states with different optimal values. To overcome such a limitation, the authors consider optimal value equivalence, wherein states are deemed equivalent if they have the same optimal value. Here again, however, problems arise: states deemed equivalent under optimal value equivalence may have markedly different MDP dynamics; in particular, they may have different optimal actions under an optimal policy and so are unsuitable for clustering. Givan et al. go on to argue that bisimulation, a refinement of the first two equivalences, is the best state equivalence for model minimization.

Bisimulation has its origins in the theory of concurrent processes (Park, 1981). Milner (1989) utilized strong bisimulation as a notion of process equivalence for his Calculus of Communicating Systems (CCS), a language used to reason about parallel processes. Bisimulation in this context can informally be seen as the largest type of matching relation, i.e. processes p and q are related iff for every a-labeled transition that process p can make to process p', process q can make an a-labeled transition to some process q' related to p', and vice versa. Alternatively, bisimulation equivalence on processes can be characterized by a modal logic known as *Hennessy-Milner logic* (Hennessy & Milner, 1985); two processes are bisimilar if and only if they satisfy precisely the same formulas.

Larsen and Skou (1991) extended this notion to a probabilistic framework. Their *probabilistic bisimulation* was developed as an equivalence notion for labeled Markov chains (LMCs). They provide characterizations of probabilistic bisimulation both in terms of a maximal matching relation and a probabilistic modal logic. The definition of bisimulation by Givan et al. is a simple extension of probabilistic bisimulation:

Definition 2.3.1. A *bisimulation relation* R is an equivalence relation on S that satisfies the following property:

$$sRs' \iff \forall a \in A, (r_s^a = r_{s'}^a \text{ and } \forall C \in S/R, P_s^a(C) = P_{s'}^a(C))$$

where $P_s^a(C) = \sum_{c \in C} P_{sc}^a$.

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

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

Bisimulation can also be formulated using fixed point theory, as has been verified for finite MDPs (Ferns, 2003). Note that the existence of a greatest fixed point in the definition below is guaranteed by the Knaster-Tarski Fixed Point Theorem, Theorem 2.1.1:

Definition 2.3.2. Let $\mathfrak{Rel}$ be the complete lattice of binary relations on S. Define $\mathcal{F} : \mathfrak{Rel} \to \mathfrak{Rel}$ by

$$s\mathcal{F}(R)s' \iff \forall a \in A, (r_s^a = r_{s'}^a \text{ and } \forall C \in S/R_{rst}, P_s^a(C) = P_{s'}^a(C))$$

where R_{rst} is the reflexive, symmetric, transitive closure of R.

Then s and s' are *bisimilar* iff $s \sim s'$ where $\sim$ is the greatest fixed point of $\mathcal{F}$.

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

Unfortunately, as an exact equivalence, bisimulation suffers from issues of instability; that is, slight numerical differences in the MDP parameters, $\{r_s^a\}$ and $\{P_{ss'}^a\}$, can lead to very different bisimulation partitions. Consider the sample MDP in Figure 2–3 with 4 states labeled x, $\hat{x}$, y, and $\hat{y}$, and 1 action labeled a. Suppose $r_{\hat{y}}^a = 0$. Then all states share the same immediate reward and transition amongst themselves with probability one. So all states are bisimilar. On the other

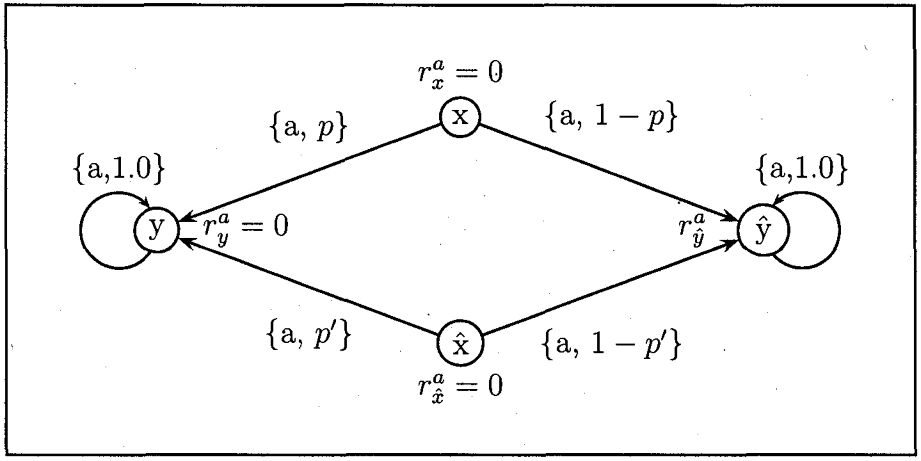


Figure 2–3: MDP demonstrating bisimulation is too brittle

hand, if $r_{\hat{y}}^a > 0$ then $\hat{y}$ is the only state in its bisimulation class since it is the only one with a positive reward. Moreover, x and $\hat{x}$ are bisimilar iff they share the same probability of transitioning to $\hat{y}$'s bisimulation class. Each is bisimilar to y iff that probability is zero. Thus, y, x, and $\hat{x}$ are not bisimilar to $\hat{y}$, $x \sim \hat{x}$ iff $p = p'$, $x \sim y$ iff $p = 1.0$, and $\hat{x} \sim y$ iff $q = 1.0$. This example demonstrates that bisimulation is simply too brittle; if $r_{\hat{y}}$ is just slightly positive, and p differs only slightly from p' then we should expect x and $\hat{x}$ to be practically bisimilar. However, an equivalence relation is too crude to capture this idea. To get around this, one generalizes the notion of bisimulation equivalence through bisimulation metrics.

2.3.1 Bisimulation Metrics

Metrics can be used to give a quantitative notion of bisimulation that is sensitive to variations in the rewards and probabilistic transitions of an MDP.

In Ferns et al. (2004; 2005) we provided the following metric generalization of bisimulation for finite MDPs:[5]

Theorem 2.3.3. *Let $c \in (0,1)$. Let* $\mathfrak{met}$ *be the space of bounded pseudometrics on S equipped with the metric induced by the uniform norm. Define $F : \mathfrak{met} \to \mathfrak{met}$ by*

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

Then :

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

Here $T_K(h)(P,Q)$ is the Kantorovich probability metric[6] applied to distributions P and Q. It is defined by the following linear program:

$$\max_{u_i} \sum_{i=1}^{|S|} (P(s_i) - Q(s_i))u_i$$

$$\text{subject to: } \forall i,j, u_i - u_j \leq h(s_i, s_j)$$

$$\forall i, 0 \leq u_i \leq \|h\|$$

[5] Results appear here in slightly modified form.

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

It can also be specified by the dual linear program

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

which can be rewritten as $\min_\lambda \mathbb{E}_\lambda[h]$ where λ is a joint probability function on $S \times S$ with projections P and Q. This discrete minimization program has an interpretation as a *Hitchcock transportation problem.* This is an instance of the minimum cost flow (MCF) network optimization problem as seen in Figure 2–4.

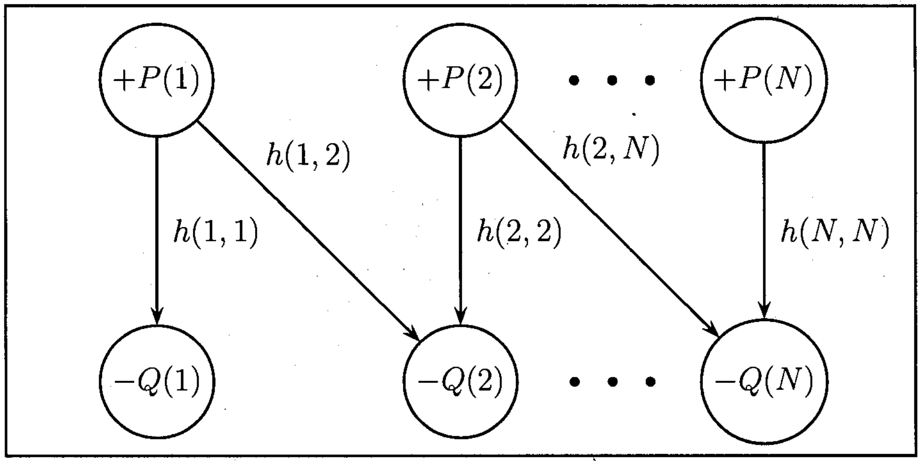


Figure 2–4: Hitchcock network transportation problem ($N = |S|$)

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

There exist strongly polynomial algorithms to compute the MCF problem (Orlin, 1988; Vygen, 2000). Therefore the Kantorovich metric in the discrete case can be computed in polynomial time, assuming of course that the state metric h is itself computable.

The key property of the Kantorovich metric is that it matches distributions, i.e. assigns them distance zero only when they agree on the equivalence classes induced by the underlying cost function. Therefore, it is not surprising that it can be used to capture the notion of bisimulation, which requires that probabilistic transitions agree on bisimulation equivalence classes. We will say more about the Kantorovich metric in the next chapter. For now, let us conclude with an example of the metric distances applied to the MDP in Figure 2–3. Using uniqueness of ρ^* and the identity $T_K(\rho^*)(\delta_x, \delta_y) = \rho^*(x, y)$ along with the fact that there is only one action, it is not hard to see that solving for $\rho*$ in the fixed point equations amounts to solving a set of linear equations. We therefore find:

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

Consider now the MDP in Figure 2–2. Even though states x and $\hat{y}$ are not bisimilar, we see that for any c they have ρ^*-distance $0.01 - 0.0095c$, which is much less than the maximum possible distance of 1; that is, they are very close to being bisimilar.

2.3.2 Value Function Bounds

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

Theorem 2.3.4 (Ferns et al., 2004). *Suppose $\gamma \leq c$. Then V^* is $\frac{1}{1-c}$-Lipschitz continuous with respect to ρ^*, i.e.*

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

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

2.3.3 Computing State-Similarity

We were able to compute the state-similarity metric by hand for the simple MDP pictured in Figure 2–3; but what can we say in the general case? In fact, the fixed point nature of the metrics permits the use of a DP algorithm in a manner analogous to the computation of the optimal value function: starting with the everywhere-zero metric, denoted by $\perp$, we iteratively apply the fixed point functional F until a desired level of accuracy is achieved. Since, as we noted, the Kantorovich operator can be computed in strongly polynomial time, we have an algorithm to calculate the state-similarity metrics - though one subject to the same shortcomings as traditional MDP dynamic programming algorithms. As only the distances are changing (and in fact converging) in the Kantorovich operator, and this object is itself an instance of an MCF LP, one immediately applicable speedup is to use cost re-optimization: that is, we can save the optimizing solutions for each Kantorovich LP between iterations and use them to begin the Kantorovich LP in the next iteration. We are thereby saving on computation time at the cost of larger space requirements. This appears slightly more promising; but, can we do better? Indeed; a promising approach to quick and efficient approximation of the distances arises from the area of statistical sampling.

Suppose P and Q are approximated using the empirical distributions P_i and Q_i. That is, we sample i points $X_1, X_2, \ldots, X_i$ independently according to P and define P_i by $P_i(x) = \frac{1}{i}\sum_{k=1}^{i} \delta_{X_k}(x)$. Similarly, write $Q_i(x) = \frac{1}{i}\sum_{k=1}^{i} \delta_{Y_k}(x)$. Then

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

where the minimum is taken over all permutations σ on i elements (see p. 12 of Villani (2002); this is a consequence of the Birkoff - von Neumann theorem). Now the Strong Law of Large Numbers (SLLN) tells us that both $\{P_i(x)\}$ and $\{Q_i(x)\}$ converge almost surely to $P(x)$ and $Q(x)$.[7] Let us write $T_K^i(h)(P,Q)$ for $T_K(h)(P_i,Q_i)$ when the empirical distributions are fixed. Then as a consequence of the SLLN, $\{T_K^i(h)(P,Q)\}$ converges to $T_K(h)(P,Q)$ almost surely; moreover replacing T_K by T_K^i in F yields a metric,

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

which converges almost surely to ρ^* as i gets large (Ferns et al., 2006).

The importance of this result stems from the fact that the expression in equation (2.3) is an instance of the assignment problem from network optimization. This is a specialized network flow problem in which the underlying network is bipartite and all flow assignments are either 0 or 1.[8] Its specialized structure allows for fast, simple solution methods. For example, the Hungarian algorithm runs in worst case time $O(i^3)$, where i is the number of samples. Still, is the resulting sampling algorithm for estimating bisimulation distances really any better than the exact algorithms?

[7] Note that both P_i and Q_i are random variables.

[8] In graph theoretic terminology, this is the problem of optimal matching in a weighted graph.

We have compared the Monte Carlo algorithm for a fixed number of samples along with the algorithms presented above, in terms of computational resources (space and time), and use in aggregation (Ferns et al., 2006). For purposes of illustration, we present here some of these results.

Experiments were run on MDPs given by an $n \times n$ grid world with two actions (move forward and rotate) and a single reward in the center of the room for $n = 3$, 5, and 7, and a flattened out version of the coffee robot MDP (Boutilier et al., 1995) in which a robot has to get coffee for a user while having to avoid getting wet. Each state in the grid world encodes both position as well as orientation of the agent; thus, the gridworld MDPs have 36, 100, and 196 states respectively. Additionally, the actions are deterministic. The coffee domain has 64 states and 4 actions, some with stochastic effects. For each domain, we computed: $\frac{1}{1-c}\rho^*$, the same with cost re-optimization, and $\frac{1}{1-c}\rho_i^*$ via sampling.

Exact computation of the Kantorovich metric in the first two methods was carried out using the MCFZIB Minimum Cost Flow solver (Frangioni & Manca, 2006). An implementation of the Hungarian algorithm for the assignment problem was used to estimate the Kantorovich distances in the third method.

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

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

Table 2–1 summarizes the running times in seconds for each method with the different discount factors. A '-' means that the algorithm failed to compute the metric.

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

Table 2–1: Running times in seconds for different metric algorithms

We also compared the amount of space used by each method. This was measured using the *massif* tool of valgrind (a tool library in Linux). Table 2–2 presents the maximum number of bytes used by each algorithm when computing the distances for each MDP; an '*' indicates an algorithm terminated prematurely due to maximum memory usage. In those cases where all algorithms were able to run to completion, the Monte Carlo algorithm either outperformed or performed comparably to the exact algorithms. Moreover, we compared the quality of the estimated distances with that of the exact distances by using each in simple aggregations schemes - and here too results were comparable (Ferns et al., 2006). All in all, when considering the tradeoff between the computational requirements of time and space, and the quality of the results, the Monte Carlo algorithm for calculating bisimulation distances significantly outperforms the others. Therefore, extending this sampling algorithm is the most promising approach to providing practical quantitative state-similarity for continuous Markov decision processes.

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

Table 2–2: Memory usage in bytes for different metric algorithms

2.4 Related Work

This work has its roots in the work of Desharnais et al. (2004) and van Breugel and Worrell (2001b). In the work of Desharnais et al. (1999; 2004) and mainly in the thesis of Desharnais (2000), the authors developed bisimulation metrics for a probabilistic transition model similar to the Markov decision process, namely the labeled Markov process (LMP) (Blute et al., 1997):

Definition 2.4.1. A *labeled Markov process* is a quadruple

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

where:

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

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

One may define bisimulation for an LMP as follows: given a relation R on S, a subset X of S is said to be R-closed if and only if $\{s' \in S | \exists s \in X.\ sRs'\} \subseteq X$. Then a *bisimulation relation* is an equivalence relation on S that satisfies the following property:

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

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

One may also define bisimulation for LMPs in terms of a modal logic: two states are bisimilar if and only if they satisfy exactly the same formulas in some fixed logic (Blute et al., 1997; Desharnais, 2000). This forms the basis for the metrics of Desharnais (2000; 1999; 2004), which are defined in terms of real-valued logical expressions. The intuition in moving to metrics is that the bisimilarity of two states is directly related to the complexity of the simplest formula that can distinguish them; the "more bisimilar" two states are, the harder it should be to find a distinguishing formula; hence, such a formula should be necessarily "big". Of course, to formalize this one needs to find some quantitative analogue of logical formulas and satisfaction. One idea of how to do this in the context of a probabilistic framework comes from (Kozen, 1983):

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

The idea is that just as the satisfaction relation maps states and propositional formulas to truth values, integration maps measures and measurable functions to extended truth values - values in the closed unit interval $[0, 1]$. On the basis of these ideas, Desharnais (2000) developed a class of logical functional expressions that could be evaluated on the state space of a given LMP to obtain values in the unit interval. A family of bisimulation metrics is then constructed by calculating the difference of these quantities for a fixed pair of states across all formulas. Formally, let $c \in (0,1]$ and let $\mathcal{F}^c$ be a family of functional expressions whose syntax is given by the following grammar:

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

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

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

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

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

Theorem 2.4.2 (Desharnais, 2000). *For every c in $(0,1]$, d^c is a 1-bounded bisimulation metric.*

In the finite case and with $c < 1$, Desharnais et al. (1999) were able to construct a decision procedure for computing the metrics to any desired accuracy; one simply replaces $\mathcal{F}^c$ in the definition above with a specially chosen finite subset of functions. However, in the general case no algorithm was provided and it remained unclear as to whether or not d^1 was computable.

Later on, van Breugel and Worrell (2001a; 2001b) worked with a slightly modified version of these metrics in a categorical setting; they used fixed point theory in conjunction with the Kantorovich probably metric to define metrics on LMPs. They were able to show that the metrics induced by the logical characterization of bisimulation and provided by Desharnais et al. coincided with their own fixed point metrics. Particularly important was their application of the Kantorovich operator and subsequent use of network linear programming to develop the first polynomial-time decision procedure for the metrics in the finite case. In recent years, these same authors have developed both a theoretical framework and decision procedure for finite LMP metrics without discounting, i.e. for $c = 1$ (van Breugel et al., 2007). Still, no work has been carried out on estimating distances for general LMPs, i.e. with infinite state spaces.

In the context of MDPs, a number of methods have been proposed for analyzing state-similarity. Li et al. (2006), for example, survey a number of state aggregation techniques for finite MDPs in an attempt to unify the theory of state abstraction: these include aggregation of states based on bisimulation, homomorphisms, value equivalence, and policy equivalence, to name a few. Müller (1997) put forth an excellently written paper containing an early sensitivity analysis result in a spirit very similar to our own; he considers abstract MDPs (with full measurable state and action spaces) in which only the stochastic transition kernels differ. He then demonstrates continuity of a sort for the optimal value function with respect to several integral probability metrics. However, these results are purely of a theoretical nature - no algorithm is provided or even suggested.

In the realm of finite MDPs, several works have analyzed the error in perturbing the parameters of a given Markov decision process. Dean et al. (1997) consider bounded-parameter MDPs, in which reward and probability parameters are specified by intervals of closed reals, and define ϵ-homogeneity: a loosening of bisimulation such that all states in the same equivalence class have reward parameters and probability parameters each differing by at most ϵ. In the paper of Even-Dar and Mansour (2003), this work was expanded upon by considering different norms on the probability parameter in the definition of ϵ-homogeneity and providing performance results specifically showing that the quality of an ϵ-homogeneous partition depended heavily on the norm in use. Most recently, Ortner (2007) has expanded upon the notion of ϵ-homogeneity in terms of *adequate* pseudometrics and used these results to analyze finite MDPs under an average reward optimality criterion.

CHAPTER 3
State-Similarity Metrics: Theory

The first thing we have to deal with in moving to infinite state spaces[1] is the issue of measurability; simply put, we can no longer specify probabilities point-to-point. One needs to look at the probabilities of sets of states, and even then, not all sets can be measured in this way. Formally, we have a potentially uncountably infinite state space, S, equipped with a sigma-algebra of measurable sets, Σ. We may think of Σ as providing some sort of "information resolution" - that is, the only pertinent sets of states are those that are measurable (and we ignore the rest). Following along these lines, we need to ensure that the reward and probability functions satisfy certain measurability conditions, that is, that they behave well with respect to measurable sets. Formally, we have the following:

3.1 Continuous Markov Decision Processes

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

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

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

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

We also make the following assumptions:

1. S is Polish space[2] equipped with its Borel sigma algebra, Σ,
2. $\sup_{s,s',a} |r_s^a - r_{s'}^a| < \infty$.
3. the image of r is contained in $[0, 1]$
4. For each $a \in A$, $r(\cdot, a)$ is continuous on S.
5. For each $a \in A$, P_s^a is (weakly) continuous as a function of s, i.e. if s_n tends to s in S then for every bounded continuous function $f : S \to \mathbb{R}$, $P_{s_n}^a(f)$ tends to $P_s^a(f)$.

3.2 Bisimulation

Our presentation of bisimulation here amounts to little more than a mild extension[3] through the addition of rewards to the definition of bisimulation given by Desharnais et al. (2004) in their work on labelled Markov processes (LMPs).

Let R be an equivalence relation on S. We now have two notions of "visibility" on S: the measurable sets, as determined by the sigma algebra on S, and the sets built up from the equivalence classes of R. Naturally, we are interested in those

[2] A topological space homeomorphic to a complete, separable metric space.

[3] In fact, this definition of bisimulation for continuous state space MDPs was first proposed to me by my supervisor, Prakash Panangaden.

sets that are visible under both criteria (measurability and equivalence). Let us formalize these concepts.

We say a set X is R-closed if the collection of all those elements of S that are reachable by R from X is itself contained in X; this is equivalent to saying that X is a union of R-equivalence classes. We write $\Sigma(R)$ for those Σ-measurable sets that are also R-closed.

Definition 3.2.1. An equivalence relation R on S is a *bisimulation relation* iff it satisfies

$$sRs' \Leftrightarrow \forall a \in A,\ r_s^a = r_{s'}^a \text{ and } \forall X \in \Sigma(R),\ P_s^a(X) = P_{s'}^a(X).$$

Bisimulation is the largest of the bisimulation relations.

Note that it is not immediately clear that bisimulation itself is a bisimulation relation (transitivity is not obvious); that this is indeed the case will be shown in the proof of theorem 3.3.2 through a fixed point characterization of bisimulation. By contrast, Desharnais et al. (2004) prove transitivity through a logical characterization of bisimulation.

3.3 Metrics

As before, we will develop state-similarity metrics over a certain space of pseudometrics on S; here, however, measurability conditions come into play.

Let $\mathfrak{Met}$ be the subset of $\mathfrak{met}$ consisting of lower semicontinuous[4] (lsc) pseudometrics. Endowing $S \times S$ with the product topology, we note that lsc functions are product measurable. Moreover, it is not hard to show that $\mathfrak{Met}$ is a closed subset of $\mathfrak{met}$, so that it is itself a complete metric space. Thus, once more we have a rich structure on our space of pseudometrics, admitting the use of important fixed point theorems, provided we construct an appropriate map on $\mathfrak{Met}$. In order to do so we first look at the best way of assigning a distance to probability measures for our purposes.

3.3.1 Probability Metrics

There are numerous ways of defining a notion of distance between probability measures on a given space (Gibbs & Su, 2002). Two typical ones are the total variation distance, capturing strong convergence of probability measures, and the Kullback-Leibler (KL) divergence,[5] capturing certain information-theoretic properties of the measures. As previously mentioned, however, the particular probability metric of which we make use is known as the Kantorovich metric. Its use in defining metrics for bisimulation was first demonstrated by van Breugel and Worrell (2001a). We present it here in greater generality; all results are taken from the books by Rachev and Rueschendorf (1998) and Villani (2002), unless otherwise stated.

[4] Recall that a function $h : S \times S \rightarrow \mathbb{R}$ is *lower semicontinuous* if whenever (s_n, s'_n) tends to (s, s'), $\liminf h(s_n, s'_n) \geq h(s, s')$.

[5] Note that the KL divergence fails to satisfy the symmetry axiom for a metric.

Given a metric $h \in \mathfrak{Met}$ and probability measures P and Q on S, the Kantorovich distance, $T_K(h)$, is defined by

$$T_K(h)(P, Q) = \sup_f (P(f) - Q(f)),$$

where the supremum is taken over all bounded measurable $f : S \to \mathbb{R}$ satisfying the Lipschitz condition: $f(x) - f(y) \leq h(x, y)$ for all $x, y \in S$. We write $Lip(h)$ for the set of all such functions.

In light of the definition of bisimulation, the importance of using the Kantorovich distance is made evident in the following lemma. Insofar as we know, this is an original result.

Lemma 3.3.1. *Let $h \in \mathfrak{Met}$. Then*

$$T_K(h)(P, Q) = 0 \Leftrightarrow P(X) = Q(X),\ \forall X \in \Sigma(Rel(h)).$$

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

$\Rightarrow$ Let $X \in \Sigma(R)$. Let $K \subseteq X$ be compact. Define $f(x) = \inf_{k \in K} h(x, k)$. Since a lsc function has a minimum on a compact set, we may write $f(x) = \min_{k \in K} h(x, k)$. In fact, f is itself lsc (Puterman, 1994, theorem B.5). Since f is measurable, $R(K) = f^{-1}(\{0\}) \in \Sigma(R)$. Now, since P is tight (as S is a complete separable metric space), $P(X) = \sup P(K)$ where the supremum is taken over all compact $K \subseteq X$. However, $K \subseteq X$ implies $K \subseteq R(K) \subseteq R(X) = X$. Since $R(K)$ is measurable, we have $P(X) = \sup P(R(K))$. Similarly, $Q(X) = \sup Q(R(K))$. Define $g_n = \max(0, 1 - nf)$. Then g_n decreases to the indicator function on $R(K)$. Also, $g_n/n \in Lip(h)$, so by assumption $P(g_n/n) = Q(g_n/n)$. Multiplying by n and taking limits gives $P(R(K)) = Q(R(K))$ and we are done. □

The Kantorovich metric arose in the study of optimal mass transportation. The following description is due to Villani (2002): assume we are given a pile of sand and a hole, occupying measurable spaces (X, Σ_X) and (Y, Σ_Y), each representing a copy of (S, Σ) (Figure 3–1). The pile of sand and the hole obviously have the same volume, and the mass of the pile is assumed to be normalized to 1. Let P and Q be measures on X and Y respectively, such that whenever $A \in \Sigma_X$ and $B \in \Sigma_Y$, $P[A]$ measures how much sand occupies A and $Q[B]$ measures how much sand can be piled into B. Suppose further that we have some measurable

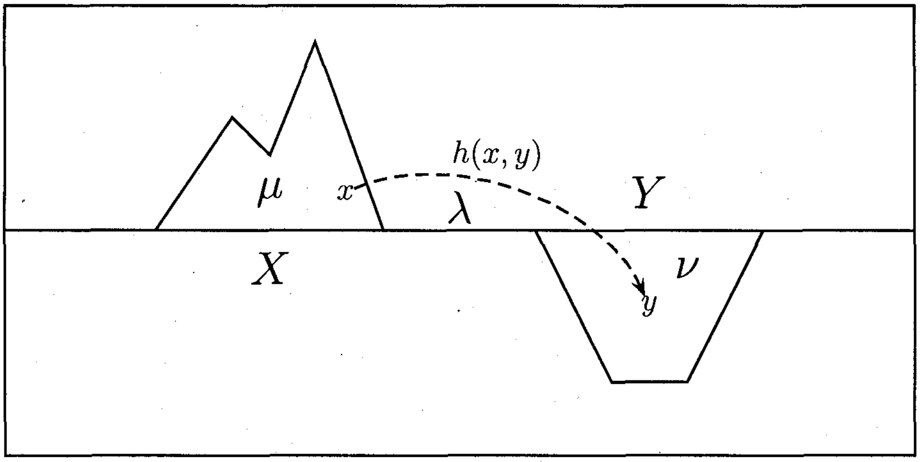


Figure 3–1: Kantorovich optimal mass transportation problem

cost function $h : X \times Y \to \mathbb{R}$, where $h(x, y)$ tells us how much it costs to transfer one unit of mass from a point $x \in X$ to a point $y \in Y$. Here we consider $h \in \mathfrak{Met}$. The goal is to determine a plan for transferring all the mass from X to Y while keeping the cost at a minimum. Such a transfer plan is modelled by a probability measure λ on $(X \times Y, \Sigma_X \otimes \Sigma_Y)$, where $d\lambda(x, y)$ measures how much mass is transferred from location x to y. Of course, for the plan to be valid we require that $\lambda[A \times Y] = P[A]$ and $\lambda[X \times B] = Q[B]$ for all measurable A and B. A plan satisfying this condition is said to have marginals P and Q, and we denote the collection of all such plans by $\Lambda(P, Q)$. We can now restate the goal formally as:

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

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

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

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

for all $(\phi, \psi) \in L^1(P) \times L^1(Q)$. As a consequence of this characterization we have the following inequality:

$$\sup_f (P(f) - Q(f)) \leq T_K(h)(P, Q) \leq \inf_{\lambda \in \Lambda(P,Q)} h(\lambda) \tag{3.1}$$

where f is restricted to the continuous functions in $Lip(h)$.

The leftmost and rightmost terms in inequality (3.1) are examples of infinite linear programs in duality. It is a highly nontrivial result that there is no duality gap in this case, as a result of the Kantorovich-Rubinstein Duality Theorem with metric cost function (Rachev & Rüschendorf, 1998, theorems 4.15 and 4.28, and example 4.24; Villani, 2002).

Note that for any h for which there is no duality gap, and for any point masses δ_x, δ_y, we have $T_K(h)(\delta_x, \delta_y) = h(x, y)$ since $\delta_{(x,y)}$ is the only measure with marginals δ_x and δ_y. As a result, we obtain that any lower semicontinuous h can be expressed as $h(x, y) = \sup_f (f(x) - f(y))$ for some family of continuous functions f (we used this property in the previous chapter to compute the state-similarity metric by hand for a very simple finite MDP).

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

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

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

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

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

Theorem 3.3.2. *Bisimulation is a closed subset of $S \times S$*

Proof. Recall that $\mathfrak{Rel}$ and $\mathfrak{Equ}$ are the complete lattices of binary relations and topologically-closed equivalence relations on S, respectively. Define $\mathcal{F} : \mathfrak{Rel} \to \mathfrak{Rel}$ by

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

Then the greatest fixed point of $\mathcal{F}$ is bisimulation.[6]

Firstly note here that R_{rst} is the reflexive,symmetric, transitive closure of R, i.e. the smallest equivalence relation containing R. Next, simply note that the fixed points of $\mathcal{F}$ are precisely the bisimulation relations. So the greatest fixed point is contained in bisimulation, and since every bisimulation relation is contained in the greatest fixed point, so is bisimulation.

Next we claim that $\mathcal{F}$ maps $\mathfrak{Equ}$ to $\mathfrak{Equ}$. That $\mathcal{F}(E)$ is an equivalence relation for a given E is obvious. To see that $\mathcal{F}(E)$ is closed, let $\{(x_n, y_n)\}$ be a sequence in $\mathcal{F}(E)$ converging to some pair of states (x, y). Let $a \in A$. By the definition of $\mathcal{F}(E)$, $r^a_{x_n} = r^a_{y_n}$ for every n. Since the reward function is continuous, taking limits yields $r^a_x = r^a_y$. Next, let ρ_E be the discrete pseudometric assigning distance 1 to two points if and only if they are *not* related by E. Since E is

[6] That $\mathcal{F}$ has any fixed points at all is a consequence of the Knaster-Tarski Theorem.

closed, ρ_E is lower semicontinuous. So the Kantorovich metric, $T_K(\rho_E)$ is well-defined. Now we can invoke the leftmost equality in (3.1) to obtain that the map $(s,s') \mapsto T_K(\rho_E)(P^a_s, P^a_{s'})$ is lsc; for since P^a_s is continuous with respect to the topology of weak convergence, $P^a_s(f)$ is continuous in the usual sense for every bounded continuous f in $Lip(\rho_E)$. So $P^a_s(f) - P^a_{s'}(f)$ is continuous on $S \times S$, and hence, lower semicontinuous. Finally, taking the supremum over all f yields that the map taking a pair of states to its Kantorovich distance with respect to ρ_E is lsc. Let X be an E-closed measurable set. Then by definition of $\mathcal{F}(E)$, $P^a_{x_n}(X) = P^a_{y_n}(X)$, which by lemma 3.3.1 means $T_K(\rho_E)(P^a_{x_n}, P^a_{y_n}) = 0$ for all n. Since $T_K(\rho_E)(P^a_s, P^a_{s'})$ is lower semicontinuous, $T_K(\rho_E)(P^a_x, P^a_y) = 0$. Again using lemma 3.3.1, $P^a_x(X) = P^a_y(X)$. So (x,y) belongs to $\mathcal{F}(E)$, i.e. $\mathcal{F}(E)$ is closed.

Now let $\sim_{\mathfrak{Equ}}$ be the least upper bound of bisimulation in $\mathfrak{Equ}$. By monotonicity, we have $\sim \, = \mathcal{F}(\sim) \subseteq \mathcal{F}(\sim_{\mathfrak{Equ}})$. So $\sim_{\mathfrak{Equ}} \subseteq \; \mathcal{F}(\sim_{\mathfrak{Equ}})$, i.e. $\sim_{\mathfrak{Equ}}$ is a postfixed point of $\mathcal{F}$; but then $\sim_{\mathfrak{Equ}} \subseteq \sim$, the latter being the greatest postfixed point.

Therefore, $\sim \, = \, \sim_{\mathfrak{Equ}}$, i.e. bisimulation is closed.

□

3.4 State Similarity Metrics

Definition 3.4.1. A pseudometric ρ on the states of an MDP is a *bisimulation metric* if it satisfies $\rho(s,s') = 0 \iff s \sim s'$.

All of the preceding theory comes together in the following crucial result. It is worth noting that our presentation is a significant extension of the work carried out by van Breugel and Worrell (2001a; 2001b) in their work on bisimulation metrics for labelled Markov processes.

Theorem 3.4.2. *Let $c \in (0,1)$ and $\mathfrak{Met}$ be the set of bounded lower semicontinuous pseudometrics on S. Define $F : \mathfrak{Met} \rightarrow \mathfrak{Met}$ by*

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

Then :

1. *F has a unique fixed point ρ^*,*
2. *ρ^* is a bisimulation metric,*
3. *for any $h_0 \in \mathfrak{Met}$, $\lim_{n\to\infty} F^n(h_0) = \rho^*$,*
4. *ρ^* is continuous on $S \times S$,*
5. *ρ^* is continuous in r and P, and*
6. *ρ^* scales with rewards.*

Before proving theorem 3.4.2, let us first make a few remarks. The first three properties tell us that a quantitative notion of bisimulation exists, and that it can be approximated. The continuity results tell us that we only need to know the metric on a dense subset, and that distances are insensitive to perturbations in the MDP parameters. The last property is not surprising, and reflects the fact that the actual numbers are not as important as the qualitative structure[7] arising from the metric.

[7] The topological or even *uniform* structures - see for example Dudley (2002) - yield the same distinguishing information with respect to bisimilarity; our specific choice of pseudometric here is influenced by theorem 3.4.10.

Lemma 3.4.3. *F has a unique fixed point ρ^*, such that for any $h_0 \in \mathfrak{Met}$,*

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

Proof. It is important to note here that we are implicitly invoking the left-most equality in (3.1) in order to correctly claim that the map taking (s, s') to $T_K(h)(P_s^a, P_{s'}^a)$ is lsc. Moreover, F is a monotone map on $\mathfrak{Met}$.

This is simply an application of the Banach Fixed Point Theorem. Here we use the dual minimization form of $T_K(\cdot)$, as given in (3.1). Note that for all $h, h' \in \mathfrak{Met}$, and for all $s, s' \in S$,

$$\begin{aligned}
F(h)(s,s') - F(h')(s,s') &\leq c \max_{a \in A}(T_K(h)(P_s^a, P_{s'}^a) - T_K(h')(P_s^a, P_{s'}^a)) \\
&\leq c \max_{a \in A}(T_K(h - h' + h')(P_s^a, P_{s'}^a) - T_K(h')(P_s^a, P_{s'}^a)) \\
&\leq c \max_{a \in A}(T_K(\|h - h'\| + h')(P_s^a, P_{s'}^a) - T_K(h')(P_s^a, P_{s'}^a)) \\
&\leq c \max_{a \in A}(\|h - h'\| + T_K(h')(P_s^a, P_{s'}^a) - T_K(h')(P_s^a, P_{s'}^a)) \\
&\leq c\|h - h'\|
\end{aligned}$$

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

The following is an original continuity result.

Lemma 3.4.4. *ρ^* is a continuous function on $S \times S$.*

Proof. Since the set of bounded continuous pseudometrics on S is a closed subset of $\mathfrak{Met}$, we need only show that F maps it to itself. So let ρ be a bounded continuous pseudometric on S. Let $a \in A$. Then continuity of r on S implies $|r_x^a - r_y^a|$

is continuous on $S \times S$. For the continuity of $T_K(\rho)(P_x^a, P_y^a)$, we appeal to the following result:

Theorem 3.4.5 (Parthasarathy, 1967). *Let X be a separable metric space and μ_n be any sequence of measures on X. Then $\mu_n \Rightarrow \mu$ if and only if*

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

for every family $\mathcal{A}_0 \subseteq C(X)$ which is equicontinuous at all the points $x \in X$ and uniformly bounded, i.e., for some constant M, $|f(x)| \leq M$ for all $x \in X$ and $f \in \mathcal{A}_0$.

This theorem implies that $T_K(\rho)$ metrizes the topology of weak convergence, provided $Lip(\rho)$ is equicontinuous and uniformly bounded. Since ρ is uniformly bounded, so is $Lip(\rho)$. As for equicontinuity at a point x, let $\epsilon > 0$. Continuity of the function $\rho(x, \cdot)$ implies that there is a neighborhood N_x of x such that for all y in N_x, $\rho(x,y) = |\rho(x,y) - \rho(x,x)| < \epsilon$. Then for any $f \in Lip(\rho)$, $|f(x) - f(y)| \leq \rho(x,y) < \epsilon$. Thus, $Lip(\rho)$ is equicontinuous. Since

$$|T_K(\rho)(P_x^a, P_y^a) - T_K(\rho)(P_{x_n}^a, P_{y_n}^a)| \leq T_K(\rho)(P_x^a, P_{x_n}^a) + T_K(\rho)(P_y^a, P_{y_n}^a)$$

we have that for any $\{(x_n, y_n)\}$ converging to (x, y), $T_K(\rho)(P_{x_n}^a, P_{y_n}^a)$ converges to $T_K(\rho)(P_x^a, P_y^a)$. Thus, continuity of $F(\rho)$ is immediate. □

As an immediate consequence of the preceding lemma, we have the following:

Corollary 3.4.6. *The topology induced by ρ^* on S is coarser than the original.*

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

Lemma 3.4.7. *ρ^* is a bisimulation metric*

Proof. It follows from lemma 3.3.1 that for any h in $\mathfrak{Met}$, $Rel(F(h)) = \mathcal{F}(Rel(h))$. Thus, $Rel(\rho^*) = \mathcal{F}(Rel(\rho^*))$ is a fixed point and so is contained in bisimulation. For the other direction, we consider the discrete bisimulation pseudometric that assigns distance 1 to pairs of non-bisimilar states; call it ρ. Since bisimulation is closed (theorem 3.3.2), ρ is lsc. So $\sim = \mathcal{F}(\sim) = \mathcal{F}(Rel(\rho)) = Rel(F(\rho))$, which implies $F(\rho) \leq \rho$. Since F is monotone, iterating F and taking limits yields $\rho^* \leq \rho$, whence it follows that $Rel(\rho^*)$ contains bisimulation.

□

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

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

Simply note that by induction $Rel(F^n(\bot)) = \sim_n$ (here, we are once again using the fact that $Rel(F(h)) = \mathcal{F}(Rel(h)))$. Since it is easily seen that

$\cap_n Rel(F^n(\bot)) = Rel(\sup_n F^n(\bot))$ and $\sup_n F^n(\bot) = \rho^*$, we have $\sim_\omega = Rel(\rho^*)$, which is bisimulation.

Thus, the nth iterate corresponds to an n-step approximation to bisimulation. Let us note that we now have three equivalent formulations of bisimulation, making this more in line with the traditional presentation of bisimulation for labeled nondeterministic transition systems: as a maximal relation, as a greatest fixed point, and as an intersection of an infinite family of equivalence relations (Milner, 1980).

Lemma 3.4.8. *If MDP M' is obtained from MDP M by setting $r' = k \cdot r$ for some scalar $k > 0$, then $\rho^*_{M'} = k \cdot \rho^*_M$.*

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

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

Lemma 3.4.9. *Let $M = (S, \Sigma, A, r, P)$ and $\widehat{M} = (S, \Sigma, A, \hat{r}, Q)$ be MDPs with common state and action spaces, and such that each satisfies the assumptions outlined at the beginning of this chapter. Let ρ and $\hat{\rho}$ be the corresponding 1-bounded bisimulation metrics given by theorem 3.4.2 with discount factor c. Then*

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

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

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

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

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

Here we have used the fact that $T_K(\varrho) = TV$ (Villani, 2002). Next, we see that

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

□

Thus, existence of the state-similarity metrics for a continuous MDP is established, along with several important properties. However, as in the finite case, perhaps the most important property of the metrics is showing that similar states have similar optimal values, and that this relation varies smoothly with similarity.

We must emphasize that in contrast with the work on LMPs, where the underlying motivation has been to analyze the validity of testing properties expressed in a modal logic on similar systems, a primary focus here is in analyzing the validity of computing optimal values (and hence, optimal policies) on similar MDPs.

3.4.1 Value Function Bounds

In moving to continuous state spaces, we must address the validity of the continuous analog of the optimality equations:

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

In general, such a V^* need not exist. Even if it does, there may not be a well-behaved, that is to say measurable, policy that is captured by it. Fortunately, there are several mild restrictions under which this is not the case; and in fact, theorem 6.2.12. of Puterman (1994) states that the optimality equations are valid provided the state space is Polish and the reward function is uniformly bounded, as is indeed the case here. Just as before, the optimal value function V^* can be expressed as the limit of a sequence of iterates V^n; we can use these to show that the optimal value function is continuous with respect to the state-similarity metrics.

Theorem 3.4.10. *Suppose $\gamma \leq c$. Then V^* is Lipschitz continuous with respect to ρ^* with Lipschitz constant $\frac{1}{1-c}$, i.e.*

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

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

We can use this result to relate the optimal values of a state and its representation in an approximant by considering the original model and its approximant as one MDP. More directly, we can use the distances themselves for aggregation with error bounds. Let us consider a simple illustration, first presented in Ferns et al. (2005), of metric-based reasoning: let $S = [0,1]$ with the usual Borel sigma-algebra, $A = \{a, b\}$, $r_s^a = 1 - s$, $r_s^b = s$, P_s^a be uniform on S, and P_s^b the point mass at s. Clearly, these MDP parameters satisfy the required assumptions.

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

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

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

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

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

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

Therefore, for $x \in B_k$,

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

In fact, we can generalize this result. Let μ be a measure on S, and $\mathcal{P}$ a finite partition of positive μ-measure. Define the μ-average finite MDP $M_{\mathcal{P}}$ by $(\mathcal{P}, A, r, P)$ where

$$r^a_B = \frac{1}{\mu(B)} \int_{x \in B} r^a_x d\mu(x)$$
$$P^a_{BB'} = \frac{1}{\mu(B)} \int_{x \in B} P^a_x(B') d\mu(x).$$

Then

$$(1-c) \cdot |V^*(s) - V^*_{M_{\mathcal{P}}}([s])| \leq \rho^*(s, [s]) \leq \frac{1}{\mu([s])} \int_{x \in [s]} \rho^*(s, x) d\mu(x),$$

where $[-]$ takes $s \in S$ to its equivalence class in $\mathcal{P}$. In other words, we can bound the distance between a state and its equivalence class by the average distance between that state and all the other states in its equivalence class. The proof of this fact is immediate, and essentially follows substitution of r and P as defined above into the fixed point equation for ρ^*.

CHAPTER 4
State-Similarity Metrics: Practice

The goal of this chapter is to develop a practical means of estimating state similarity in a continuous MDP using the theory of bisimulation metrics developed in the previous chapter. In practice, though dealing with infinite state spaces, we need to get our hands on some finite structure with which we can work; it is for this reason that we will restrict our attention to those Markov decision processes with compact metric state spaces.

4.1 Distance Approximation Schemes

Based on a comparison of distance-estimation schemes for finite MDPs (Ferns et al., 2006), the most promising method for estimating bisimulation distances for continuous MDPs would appear to be the sampling method: one samples all probability mass functions, replaces each with an empirical distribution built from the resulting samples, and repeatedly applies the fixed point bisimulation functional to the new MDP. Supposing for the moment that one can enumerate and sample from a compact metric space with full-fledged probability measures, the only problem in this procedure is the validity of replacing the original MDP with the sampled version. In other words, if we replace the probability measures in our MDP with empirical measures, is it still true that the bisimulation metric on the sampled MDP will converge to the true bisimulation metric as the number of samples increases?

Fortunately, with some minor modifications the answer is yes. In order to prove this, we will need to make use of a uniform Glivenko-Cantelli theorem. Such theorems typically characterize uniform convergence of empiricals to means, and are ubiquitous throughout machine learning (Anthony, 2002). Formally, one says that a family of real-valued measurable functions $\mathcal{F}$ is a *strong uniform Glivenko-Cantelli class* if and only if

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

where the outermost supremum is taken over all probability measures on the state space, $\mathbb{P}$ is the underlying sampling probability measure and μ_m is the empirical measure of μ on m samples.[1] Let us take a moment to consider what this means in the context of the Kantorovich distances. Suppose *Lip*(h) is a uniform Glivenko-Cantelli class for pseudometric h. Then the uniform Glivenko-Cantelli property tells us that $T_K(h)(\mu, \mu_i)$ converges to zero $\mathbb{P}$-almost surely for all μ *and* this convergence is uniform over all μ.

The question as to which classes constitute uniform Glivenko-Cantelli classes and under what conditions is an important area of active research. Recall that *Lip*(h) refers to the set of bounded measurable functions on S that are 1-Lipschitz continuous with respect to h (Section 3.3.1). Recall too, however, that the value of

[1] Recall that each μ_m is actually a random variable over some ambient probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We use the outer probability $\mathbb{P}^*$ in place of $\mathbb{P}$ in the definition of uniform Glivenko-Cantelli class to avoid issues of measurability.

the Kantorovich metric is unchanged if we restrict attention to only the continuous functions in *Lip*(h). Henceforth, *Lip*(h) will refer to the set of bounded measurable functions that are continuous on S and 1-Lipschitz continuous with respect to h.

Lemma 4.1.1. *Lip*(ρ^*) *is a uniform Glivenko-Cantelli class.*

Proof. Since S is assumed to be a compact metric space, S equipped with ρ^* is again a compact (pseudo)metric space with a coarser topology (corollary 3.4.6); hence, (S, ρ^*) is totally bounded. Our result then is contained in the proof of the following proposition:

Proposition 4.1.2 (Dudley et al., 1991). *For any separable metric space* (X, d) *and* $0 < M < \infty$, $\widetilde{F}_M := \{f : \|f\|_{BL} \leq M\}$ *is a universal Glivenko-Cantelli class. It is a uniform Glivenko-Cantelli class if and only if* (X, d) *is totally bounded.*

A few remarks are warranted in regard to the application of this proposition to Lemma 4.1.1. Firstly, the proof of Proposition 4.1.2 in Dudley et al. (1991) shows us that the implication that $\widetilde{F}_M$ is a uniform Glivenko-Cantelli class provided (X, d) is totally bounded depends solely on the integral probability metric induced by $\widetilde{F}_M$ metrizing weak convergence and $\widetilde{F}_M$ itself being totally bounded in uniform norm - and this remains valid in the context of pseudometric spaces. In other words, if we can show that the integral probability metric induced by *Lip*(ρ^*), $T_K(\rho^*)$, metrizes weak convergence and that *Lip*(ρ^*) is totally bounded, then it will follow that *Lip*(ρ^*) is uniform Glivenko-Cantelli.

Theorem 3.4.5 shows us that $T_K(\rho^*)$ metrizes weak convergence (as outlined in the proof of lemma 3.4.4); so we need only show that *Lip*(ρ^*) is totally bounded.

Since (S, ρ^*) is totally bounded, so is $\widetilde{F}_2$. Now since ρ^* is continuous on $S \times S$, $Lip(\rho^*)$ can simply be defined as those functions that are 1-Lipschitz with respect to ρ^*; for any such function f we have

$$||f|| \leq ||\rho^*|| \leq 1, \text{ and}$$

$$||f||_L := \sup\left\{ \frac{|f(x) - f(y)|}{\rho^*(x,y)} : x, y \in X \text{ such that } \rho^*(x,y) \neq 0 \right\} \leq 1.$$

Therefore, $||f||_{BL} := ||f|| + ||f||_L \leq 2$, whence it follows that $Lip(\rho^*) \subseteq \widetilde{F}_2$ (here $|| \cdot ||_L$ and $|| \cdot ||_{BL}$ are the Lipschitz and bounded-Lipschitz norms, respectively). As a subset of a totally bounded set is itself totally bounded, we are done.

□

How does this help us? Recall that as a first step in our distance approximation scheme, we would like to replace each probability measure on the space with an empirical measure and use Theorem 3.4.2 to guarantee existence of bisimulation metrics. However, in order to use that we require the map taking states to empirical measures to be continuous - and in general this need not be the case. We may circumvent this issue by replacing the Kantorovich operator with one that is defined on *all* real-valued functions, not just the measurable ones. For a fixed i, define for empiricals $\mu_i = \frac{1}{i}\sum_{j=1}^{i} \delta_{X_j}$ and $\nu_i = \frac{1}{i}\sum_{j=1}^{i} \delta_{Y_j}$ and bounded-pseudometric h,

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

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

Proposition 4.1.3. *Let $c \in (0,1)$ and $i \in \mathbb{N}$. Define F_i : $\mathfrak{met} \rightarrow \mathfrak{met}$ by*

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

Then :

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

Thus, the proposed statistical estimates $\{\rho_i^*\}$ to ρ^* exist; yet, how do we know that they actually converge to ρ^*? It is not hard to see that

$$\|\rho_i^* - \rho^*\| \leq \frac{2c}{1-c} \sup_{a\in A, s\in S} T_K(\rho^*)(P_{i,s}^a, P_s^a). \tag{4.1}$$

Simply note that

$$\begin{aligned} |\rho_i^*(s,s') - \rho^*(s,s')| &\leq c\max_{a\in A}|T_K^i(\rho_i^*)(P_{i,s}^a, P_{i,s'}^a) - T_K(\rho^*)(P_s^a, P_{s'}^a)| \\ &\leq c\max_{a\in A}|T_K^i(\rho_i^*)(P_{i,s}^a, P_{i,s'}^a) - T_K^i(\rho^*)(P_{i,s}^a, P_{i,s'}^a)| \\ &\quad + c\max_{a\in A}|T_K(\rho^*)(P_{i,s}^a, P_{i,s'}^a) - T_K(\rho^*)(P_s^a, P_{s'}^a)| \end{aligned}$$

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

$$\leq c\|\rho_i^* - \rho^*\| + c\max_{a\in A}(T_K(\rho^*)(P_{i,s}^a, P_s^a) + T_K(\rho^*)(P_{i,s'}^a, P_{s'}^a))$$

$$\leq c\|\rho_i^* - \rho^*\| + 2c\sup_{a,s} T_K(\rho^*)(P_{i,s}^a, P_s^a)$$

and the result follows. This is where the uniform Glivenko-Cantelli property comes into play: we would like to use it to show that the quantity on the right-hand side of inequality 4.1 tends to zero almost surely. Unfortunately, we face a problem in the form of the supremum over the possibly uncountably infinite set S. While the uniform Glivenko-Cantelli theorem indeed tells us that empiricals converge in Kantorovich distance to their measure almost surely for each measure, and even over all measures almost surely for a countable set of measures, it does not dictate that all measures converge at the same rate uniformly almost surely. Here compactness comes to the rescue.

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

$$\max_{a\in A} T_K(\rho^*)(P_x^a, P_y^a) \leq \frac{1}{c}\rho^*(x, y) < \frac{\epsilon}{c}.$$

Let $[-] : S \to U$ be a mapping such that $d(s, [s]) < \delta_c(\epsilon)$ for every s in S and the image $[S]$ is finite; that this can be done is a consequence of U being dense in S and S being compact. Next, if $\mu_i = \frac{1}{i}\sum_{j=1}^i \delta_{X_j}$, define $[\mu_i]$ to be $\frac{1}{i}\sum_{j=1}^i \delta_{[X_j]}$.

Then for any μ_i

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

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

Theorem 4.1.4. *Let $c \in (0,1), i \in \mathbb{N}$, and $\epsilon > 0$. Define $F_{i,\epsilon} :$ $\mathfrak{met} \longrightarrow \mathfrak{met}$ by*

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

Then :

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

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

$$\|\rho^*_{i,\epsilon} - \rho^*\| \leq \frac{1}{1-c}(6\epsilon + 2c \max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)). \tag{4.2}$$

As in the previous proposition, let us note that

$$|d^*_{i,\epsilon}(s,s') - d^*(s,s')| \leq (1-c)\max_{a\in A}(|r^a_{[s]} - r^a_{[s']}| - |r^a_s - r^a_{s'}|)$$

$$+ +c\max_{a\in A}|T^i_K(\rho^*_{i,\epsilon})([P^a_{i,[s]}], [P^a_{i,[s']}]) - T_K(\rho^*)(P^a_s, P^a_{s'})|$$

$$
\begin{aligned}
&\leq (1-c)\max_{a\in A}|r^a_{[s]} - r^a_s| + \max_{a\in A}|r^a_{[s']} - r^a_{s'}| \\
&\quad + c\max_{a\in A}|T^i_K(\rho^*_{i,\epsilon})([P^a_{i,[s]}],[P^a_{i,[s']}]) - T^i_K(\rho^*)([P^a_{i,[s]}],[P^a_{i,[s']}])| \\
&\quad + c\max_{a\in A}|T_K(\rho^*)([P^a_{i,[s]}],[P^a_{i,[s']}]) - T_K(\rho^*)(P^a_s, P^a_{s'})| \\
&\leq \rho^*(s,[s]) + \rho^*(s',[s']) + c\|\rho^*_{i,\epsilon} - \rho^*\| \\
&\quad + c\max_{a\in A}\{T_K(\rho^*)([P^a_{i,[s]}], P^a_{i,[s]}) + T_K(\rho^*)(P^a_{i,[s]}, P^a_{[s]}) + T_K(\rho^*)(P^a_{[s]}, P^a_s) \\
&\quad + T_K(\rho^*)(P^a_{s'}, P^a_{[s']}) + T_K(\rho^*)(P^a_{[s']}, P^a_{i,[s']}) + T_K(\rho^*)(P^a_{i,[s']}, [P^a_{i,[s']}])\} \\
&\leq c\|\rho^*_{i,\epsilon} - \rho^*\| + 6\epsilon + 2c\max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)
\end{aligned}
$$

and the bound follows immediately.

By the Uniform Glivenko-Cantelli property, the rightmost term in inequality 4.2 tends to zero $\mathbb{P}$-almost surely (incidentally, dependent on ϵ); for, given a finite set $\mathcal{U}$ of measures, we have for a given $\varepsilon > 0$

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

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

$$
\lim_{i\to\infty}\|\rho^*_{i,\epsilon} - \rho^*\| \leq \frac{6\epsilon}{1-c} \tag{4.3}
$$

except for a set N_ϵ of $\mathbb{P}$-measure zero. Consider now only rational ϵ, and let N be the union of the collection $\{N_\epsilon\}$ over all such ϵ. Then save for N, inequality 4.3 holds for every ϵ, and N has $\mathbb{P}$-measure zero. So letting ϵ tend to zero in the same inequality, we find that $\rho_{i,\epsilon}$ converges to ρ^*, as was to be shown. □

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

4.2 On the Road to an Algorithm: the Question of Representation

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

How does one represent a probability measure on a continuous space? In the discrete setting, one of two approaches traditionally suffices: either probabilities can be specified point-to-point in a probability matrix, or one restricts attention to a parameterized class of probability mass functions. This latter approach also applies to Euclidean spaces, where one typically works with probability density functions. Although one may argue that both approaches can be suitably extended in the setting of a compact metric space (the interested reader is directed to the works of Edalat (1997) and Webster (2006)) we will focus on the approach taken by Bouchard-Côté et al. (2005).

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

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

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

- r is a computable function from $U \times A$ to $[0, 1]$

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

If a means of enumerating an ϵ-net for a given problem does not make itself obvious, then, as noted by Clarkson (2006), an ϵ-net X can be constructed by the following greedy algorithm essentially devised by Gonzalez (1985): given input $\epsilon > 0$ and maximum allowable ϵ-net size k, pick $s \in U$ arbitrarily, and set $X := \{s\}$. Then repeat the following: pick an $s \in U - X$ that maximizes $d(s, X) = \min\{d(s, x) : x \in X\}$. If $d(s, X) < \epsilon$ or $|X| \geq k$ then stop; otherwise, set $X := X \cup \{s\}$, and continue. Then X is an ϵ'-net for some $\epsilon' \leq \epsilon$ *provided* k is large enough; specifically, $\epsilon' := d(s, X - \{s\})$ where s was the last state to be added to X. The only problem in immediately applying this algorithm to the general case is in finding the element s in U that maximizes $d(s, X)$. We can get around this by sampling according to μ; either sample the space beforehand and run the algorithm on the finite samples, or as suggested in Bouchard-Côté et al. (2005) replace the maximum with the essential-supremum with respect to μ and approximate this via sampling according to μ and maximizing over the samples. The resulting heuristic should provide an ample covering of U.

Our algorithm then is as follows: given a positive rational epsilon, enumerate a $\delta_c(\epsilon)$ cover X. Define $[s]$ to be the nearest neighbor of s in X according to d. Sample the probability distributions induced by X and use $[-]$ to restrict them to X. Finally, perform the iteration algorithm on X, as in the finite case. Figure 4–1 provides pseudocode for estimating distances to within an iteration error of δ for a given ϵ and ϵ-net X.

4.3 Estimation Error

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

$$\begin{aligned}
&\|(F_{i,\epsilon})^n(\bot) - \rho^*\| \leq \|(F_{i,\epsilon})^n(\bot) - \rho^*_{i,\epsilon}\| + \|\rho^*_{i,\epsilon} - \rho^*\| \\
&\leq \frac{c^n}{1-c}\|F_{i,\epsilon}(\bot)\| + \|\rho^*_{i,\epsilon} - \rho^*\| \\
&\leq \frac{c^n}{1-c}(1-c)\cdot(1-c)\frac{1}{1-c}(6\epsilon + 2c \max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)) \\
&\leq c^n + \frac{6\epsilon}{1-c} + \frac{2c}{1-c} \max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)
\end{aligned}$$

Let $\varepsilon_\sim$, $\varepsilon_{[-]}$, and $\varepsilon_{\mathbb{P}}$ denote c^n, $\frac{6\epsilon}{1-c}$, and $\frac{2c}{1-c}\max_{a\in A, u\in[S]} T_K(\rho^*)(P^a_{i,u}, P^a_u)$; these are, respectively, the bisimulation, discretization, and sampling errors. In the next few sections, we will try to bound these to within some prescribed degree of accuracy.

4.3.1 Bisimulation Error

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

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

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

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

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

Figure 4–1: Pseudocode for estimating bisimulation distances

4.3.2 Discretization Error

In some sense, bounding the discretization error is hopeless - we need to know how ρ^* varies with d and in general, this is information that we just do not have. However, there is some hope; recall that what we need is some way of specifying a

$\delta_c(\epsilon)$ so that $d(x, y) < \delta_c(\epsilon)$ implies $\rho^*(x, y) < \epsilon$. Suppose we can bound ρ^* from above by a continuous metric m; define the modified metric d_m to be $\max(d, m)$. Then, as $d \leq d_m$ and d_m is continuous with respect to d, we have that d_m and d are compatible metrics, i.e. they induce the same topology on S. Therefore, we could use d_m in place of d and simply take $\delta_c(\epsilon)$ to be ϵ; but how do we find m? More to the point - as ρ^* is itself a candidate - how do we find an m that is easier to compute than ρ^*?

We propose here a heuristic for computing such an m. We cannot hope to bound the discretization error in computing m due to the reasons mentioned above; however, we hope to shift the focus of the discretization error onto how r and P vary with d. In other words, if we discretize the state space using an ϵ-net with respect to d_m then we will be able to set $\varepsilon_{[-]} = \frac{6\epsilon}{1-c} + \varepsilon_m$ where ε_m, the estimation error for d_m, hopefully varies much more closely with d than does ρ^*.

Let $\mathfrak{Met}_C \subseteq \mathfrak{Met}$ denote the space of bounded continuous pseudometrics on S. Define $R \in \mathfrak{Met}_C$ by

$$R(x, y) = \max_{a \in A} |r_x^a - r_y^a|$$

and the operator $T : \mathfrak{Met}_C \longrightarrow \mathfrak{Met}_C$ by

$$T(h)(x, y) = \max_{a \in A} (P_x^a \otimes P_y^a)(h),$$

where $\mu \otimes \nu$ is the product measure of μ and ν. The fact that $T(h)$ is symmetric follows from the Fubini-Tonelli Theorem (see for example Folland (1999)), which allows one to change the order of integration in an iterated integral. The fact that $T(h)$ is continuous for h in $\mathfrak{Met}_C$ follows from the fact that for separable metric spaces the limit of the product of weakly converging measures is the product of the limits of those measures, i.e. if $\mu_n \Rightarrow \mu$ and $\nu_n \Rightarrow \nu$ then $\mu_n \otimes \nu_n \Rightarrow \mu \otimes \nu$ (Billingsley, 1968). We immediately have that for any $h \in \mathfrak{Met}_C$,

$$F(h) \leq (1-c)R + cT(h),$$

where F is the fixed point operator for ρ^*. Finally, we define

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

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

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

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

Let us assume that we can compute $(P_x^a \otimes P_y^a)(h)$ for any computable h, e.g. through numerical integration, sampling, etc. Then we can compute m for any pair of states by iterating T until c^n is less than some prescribed degree of accuracy and computing the nth partial sum.[3] Finally, d_m can be computed as the maximum of m and d, and can even be taken to be 1-bounded.[4]

4.3.3 Sampling Error

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

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

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

$$\sup_{u \in [S]} \mathbb{P}^*\big(T_K(\rho^*)(P^a_{i,u}, P^a_u) > \frac{1-c}{2c}\Delta\big) \leq \frac{\alpha}{|A||[S]|}. \tag{4.4}$$

The lower bound on the number of samples required to achieve the specified level of accuracy with the specified probability is known as the *sample complexity*. A large number of bounds exist in terms of various notions of dimension, e.g.

[3] This, of course, introduces the additional estimation error ε_m.

[4] We may replace d with the compatible 1-bounded metric $\frac{d}{1+d}$.

VC-dimension, the fat-shattering dimension, covering numbers (Anthony, 2002); in general, a specific bound will depend on the structure of the metric space in question. As such, we are not able to provide specific bounds for the sample complexity in full generality. However, as an example, the following asymptotic lower bound for 4.4 can be obtained from Theorem 3.6 of Alon et al. (1997):

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

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

4.3.4 Computational Complexity

Precise computational complexity results are difficult to come by due to the application of this work to general metric spaces. The particular performance will depend on the structure of a given space - and this in turn can be represented by a number of proposed measures of metric space dimension (Clarkson, 2006). However, the previous sections do give an idea of the space and time requirements in computing distances to a given level of accuracy with a given probability. A quick glance will tell us that it would be very expensive to attempt to compute distances to within a very small degree of error with high probability - but this is none too surprising. Previous work (Even-Dar & Mansour, 2003) has shown that finding the minimal ϵ-equivalent MDP for a given finite MDP in tabular form is NP-hard and computing the bisimulation metrics obviously solves that problem. In practice we fix the number of samples in our sampling procedure and sacrifice accuracy for improved running times, e.g. for a fixed number of samples i and a given discretization $[-]$, let n be the number of discretized states in $[S]$ and m be the number of actions; then computing the state-similarity distances to within a bisimulation error of δ requires time $O(\frac{\ln\delta}{\ln c}mn^2i^3)$. In order to see this, let us refer to the pseudocode in Figure 4–1: in the initialization phase, for every discrete state and for every action, a sample is obtained and a nearest neighbour search is peformed, i times. let us assume that sampling takes constant time; then this requires $O(nmi(O(1)+n))$, or $O(mn^2i)$ steps. In the algorithm's main loop, we iterate the following procedure for $\lceil\frac{\ln\delta}{\ln c}\rceil$ steps: for every pair of states and for every action, peform the Hungarian algorthim on their induced empirical probability distributions, taking $O(i^3)$ steps for each pair and leading to a total of $O(n^2mi^3)$ steps. Then for every pair of states a maximization must be performed over the m actions, requiring a total of $O(n^2m)$ steps. So the main loop requires $O(\frac{\ln\delta}{\ln c}(mn^2i^3+mn^2))$, or $O(\frac{\ln\delta}{\ln c}mn^2i^3)$ steps. The entire algorithm then requires $O(mn^2i)+O(\frac{\ln\delta}{\ln c}mn^2i^3)=O(\frac{\ln\delta}{\ln c}mn^2i^3)$ steps. Future algorithmic efficiency, however, will require the imposition of several structural/representational conditions and learning just how to exploit these.

CHAPTER 5
Experiments

In this chapter we perform a few illustrative experiments to demonstrate the use of our distance-approximation scheme in practice. Software was written using the Java programming language and Java's pseudorandom number generator was used to sample states. Experimental data was analyzed in MATLAB. The distance-approximation algorithm was coded by myself, using an implementation of the Hungarian Algorithm by Konstantinos A. Nedas (2005).

5.1 A Simple MDP

For our initial experiment we used the simple MDP described in section 3.4.1: $S = [0, 1]$ with the usual Borel sigma-algebra, $A = \{a, b\}$, $r_s^a = 1 - s$, $r_s^b = s$, P_s^a is uniform on S, and P_s^b is the point mass at s. The simple structure of this MDP allows us to compute exactly the form of the bisimulation metric: $\rho^*(x, y) = |x - y|$. Therefore, we can use this expression to test the accuracy of our algorithm as we vary the level of discretization and the number of samples for a fixed bisimulation error.

For our experiments, we considered metric discount factors in $\{0.1, 0.5, 0.9\}$ and discretized the unit interval by dividing it into subintervals of size ϵ, for ϵ belonging to $\{0.025, 0.050, 0.075, \ldots, 0.450, 0.475, 0.500\}$, and choosing the left endpoints of each subinterval to be in our ϵ-net, e.g. for $\epsilon = 0.250$ the unit interval $[0, 1]$ is divided into $\{[0, 0.250), [0.250, 0.500), [0.500, 0.750), [0.750, 1]\}$, which gives the ϵ-net $\{0, 0.250, 0.500, 0.750\}$. For constructing empirical measures, we let the number of samples vary from 1 to 30. Throughout, we kept the bisimulation error fixed at the low value of 0.001. For each setting of the parameters, we computed the metric estimate $\hat{\rho}$, its computation time in milliseconds, and its distance from the exact metric with respect to the uniform norm, i.e.

$$\|\hat{\rho} - \rho^*\| = \sup_{x,y \in [0,1]} |\hat{\rho}(x,y) - \rho^*(x,y)|.$$

For estimating this latter term we used an ϵ-discretization of $[0,1]$ with $\epsilon = 0.01$ and used a nearest-neighbour mapping to extend $\hat{\rho}$ to this setting. For each setting, we performed these calculations for thirty independent runs. The results, averaged over the thirty runs, are depicted in Figure 5–1 and Figure 5–2.

The timing results coincide with what we would expect given the discussion at the end of the previous chapter, e.g. for a fixed discretization ϵ, there appears to be an order of growth of $O(n^3)$ in terms of the number of samples. The increase in distance-approximation error with the discount factor on the other hand is at first glance unsettling: after all, the function we are attempting to estimate, $\rho^*(x,y) = |x - y|$, is independent of the metric discount factor c!

In fact, this too is to be expected: the linear increase in error with the discretization can be attributed solely to the use of the nearest-neighbor method in calculating the approximation error; as for the sharp increase in the approximation error in the case of higher metric discount factors, low discretization, and smaller sample sizes, this only serves to illustrate the fact that higher values of c correspond to a greater emphasis on future distances in the recursive definition of ρ^*.

In other words, since we require many more iterations for higher values of c, the sampling and discretization error build up to a greater extent (see, for example, the error bounds in section 4.3.3). Since high values of c are common in practical application, this at first appears to be troubling; notice however, that the data for our simple example shows that for a fixed discretization, the error decreases sharply with the number of samples. So one can obtain a good approximation even for higher values of c, simply by increasing the number of samples.

5.2 Puddle World

In this section we investigate a more realistic problem through a variant of "Puddle World" (Sutton, 1996), as pictured in Figure 5–3. In this problem, an agent moves throughout the unit square $[0, 1] \times [0, 1]$ according to one of four possible actions: move up, down, left, or right by 0.05, up to the limits of the space. For each action, there is a stochastic noise in the form of a Gaussian with mean zero and standard deviation 0.025 (either vertical or horizontal, depending on the direction of the chosen action). There are also puddles, circles of radius 0.1 whose centers belong between $(0.1, 0.75)$ and $(0.45, 0.75)$, and $(0.45, 0.4)$ and $(0.45, 0.75)$, as well as a goal area consisting of all those positions above the line $x + y = 1.9$. Rewards are assigned according to position: a reward of $0.4 - 4d_1$ is achieved for positions within the puddles, where d_1 is the straight-line distance into the puddle, a reward of $0.4 + 6\sqrt{2}d_2$ is achieved for positions within the goal area, where d_2 is the straight-line distance past the goal line, and a reward of 0.4 is achieved everywhere else. Note that we have modified the reward function of the original problem in order to meet the continuity and boundedness conditions set

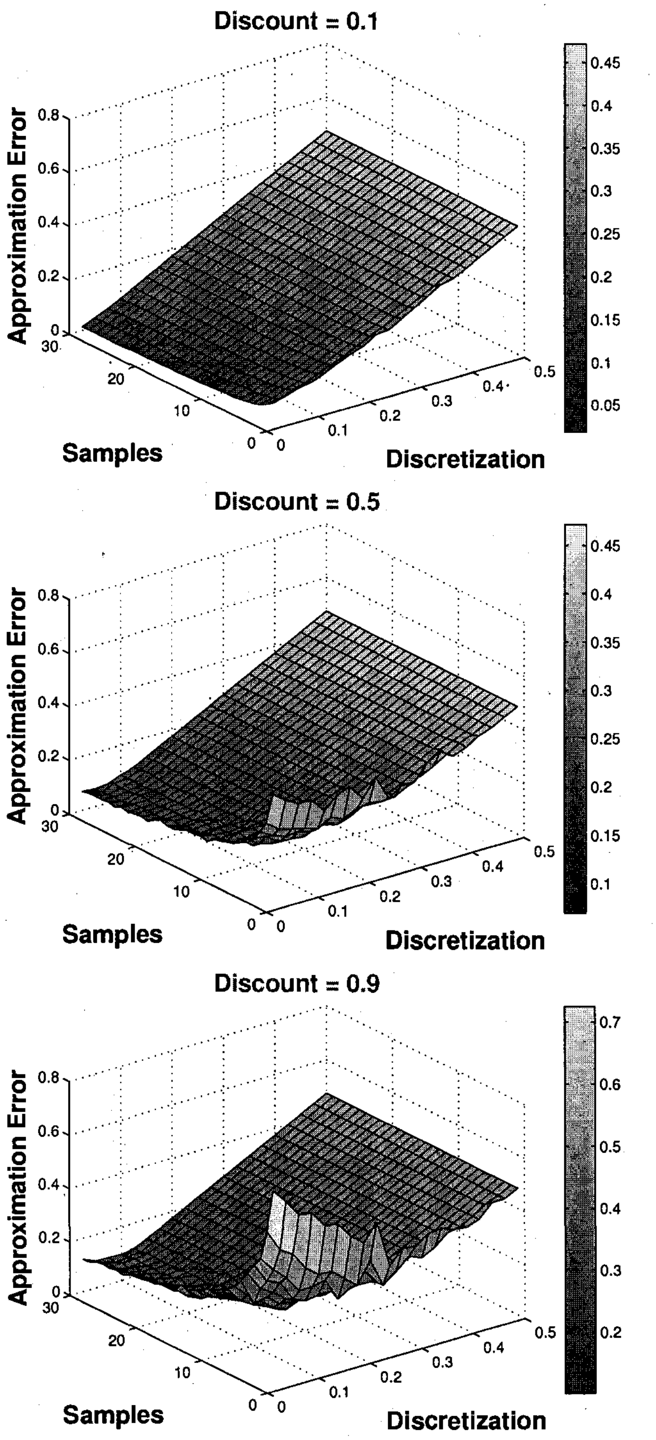
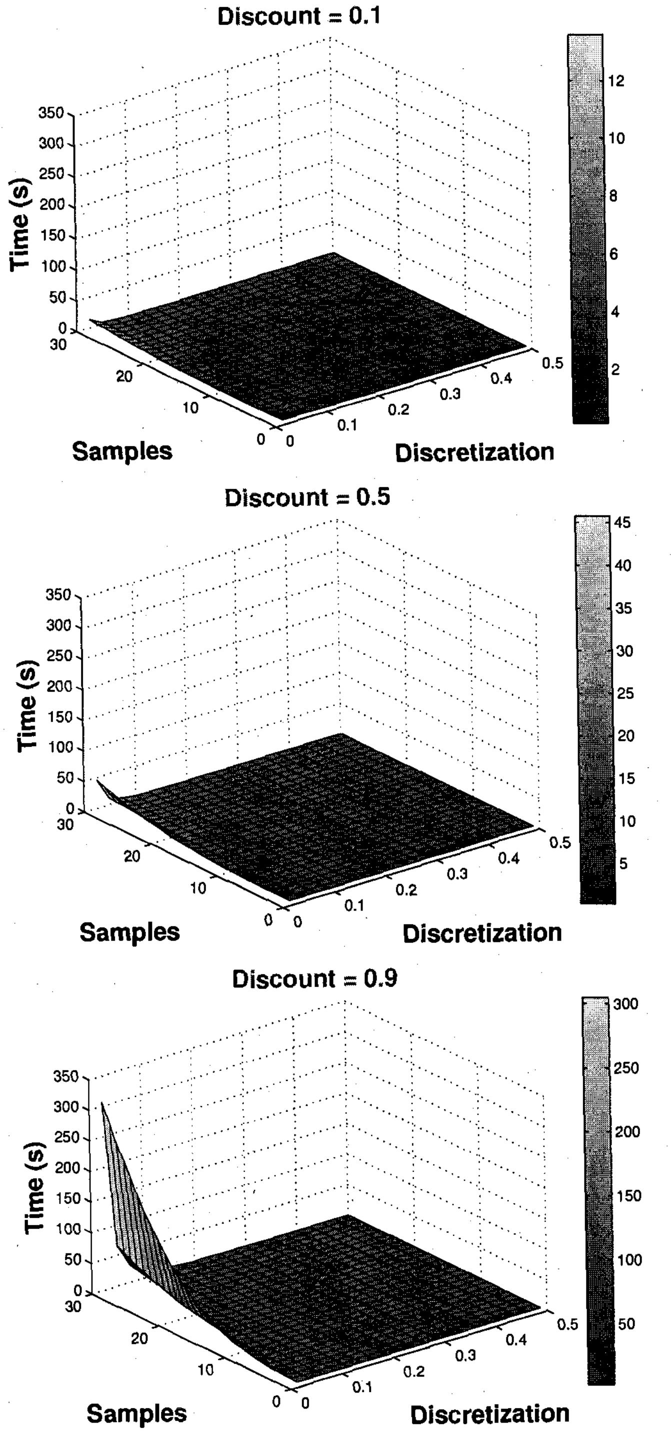


Figure 5–1: Approximation error for estimated metrics



Discount = 0.1

Discount = 0.5

Discount = 0.9

Figure 5–2: Computation time for estimated metrics

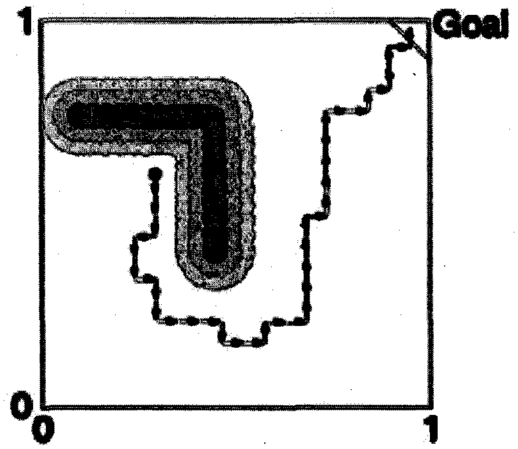
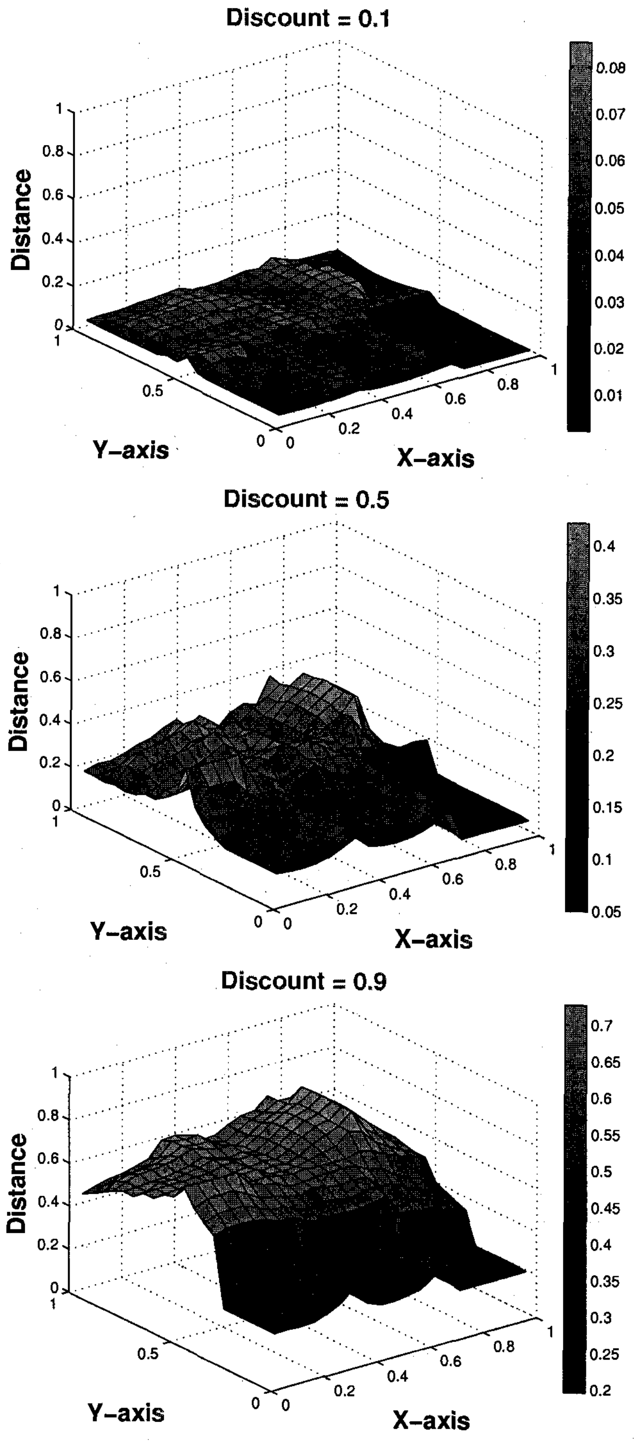


Figure 5–3: An agent moving throughout a modified Puddle World

out at the beginning of Chapter 3. The goal of this Markov decision problem is to learn a policy that would get and keep the agent into the goal area while avoiding the puddles.

Here, we consider the effects of using a deterministic model of the environment in place of the original. Our deterministic model is obtained by simply neglecting the Gaussian noise. We then estimate the bisimulation distance between a state, or position, in the original model and the corresponding position state in the deterministic model. An overall small magnitude in distance would justify the use of the more easily solvable deterministic model.

For this experiment, we again considered metric discount factors in $\{0.1, 0.5, 0.9\}$ and discretized the unit interval into subintervals of size 0.05, choosing the left endpoints to represent the subintervals. We then took the product of this set of points with itself in order to obtain a grid of points that covered the unit square.



Discount = 0.1

Discount = 0.5

Discount = 0.9

Figure 5–4: Puddle World distances

Empirical measures were obtained by sampling 10 states each. We then estimated the distances between each state and its deterministic copy, once more using a bisimulation error of 0.001. We performed this calculation for thirty independent runs for each setting of the discount factor. Results were averaged over the thirty independent runs and are pictured in Figure 5–4.

For our Puddle World problem domain, these results imply that one can justify use of a solely deterministic model for low metric discount factors (and hence low value discount factors), as the distances are indeed everywhere negligible for $c = 0.1$, and very small, being roughly no more than 40% of the maximum distance in small areas of the domain, for $c = 0.5$. For the high discount factor of $c = 0.9$ on the other hand, we once more see that states are distinguished to a greater extent as we look further ahead into the future. Here the use of a deterministic model can only be justified in roughly the bottom half of the physical domain. This coincides with what we would expect, since states closer to areas of greater reward differential, i.e. the puddles and the goal zone, carry a greater risk of being distinguished by those rewards.

CHAPTER 6
Conclusions

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

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

6.1 Future Work

There are many interesting directions possible for future investigation. Chief among these is the question of whether or not the results appearing in this work remain valid with less stringent or alternative conditions on the Markov decision problem parameters. Let us make a few quick remarks on this matter: firstly, the work of Desharnais (2000) for LMPs provides ample evidence that existence of our metrics should remain valid in at least analytic spaces. Following along the lines of Müller (1997), we may replace uniform boundedness of rewards with boundedness in terms of a bounding "weight" function, which controls the rate at which the functions grow - this essentially amounts to replacing all uniform norms by weighted uniform norms in the proofs of this work. Promising work on Kantorovich duality (Dedecker et al., 2004) may allow us to show that the mapping of states to the Kantorovich distance of their induced distributions in Theorem 3.4.2 is a measurable mapping, thereby allowing us to remove continuity conditions on the reward and probability parameters, at least in existence proofs.

There are problem instances where each time step is equally important, and discounting is unsuitable; in these cases an average reward optimality criterion (Puterman, 1994) is preferable for finding optimal polices for a given Markov decision process. We conjecture that $\lim_{c \to 1} \rho^*$ may yield a bisimulation metric suitable for analyzing average reward Markov decision problems.

We could also consider applying our work to extensions of bisimulation. Desharnais et al. (2002), for example, utilize *weak bisimulation* instead of bisimulation when developing a quantitative notion of state-similarity for a finite probabilistic transition system: essentially, states are deemed equivalent if they match over a sequence of transitions, rather than precisely at every step.

An immediate concern is that the algorithm proposed in this work was tested merely to illustrate its validity; a more extensive empirical investigation needs to be performed and will be carried out at a later stage. In practice, however, MDPs are rarely represented explicitly; instead, researchers usually work with factored representations (Boutilier et al., 2000), wherein the state space is represented by a family of state variables. Each MDP parameter is then compactly represented in terms of these variables, e.g. using dynamic Bayes nets or multi-terminal binary decision diagrams, yielding a compact representation of an MDP. If metric calculation can be adapted to work solely with the factored representation, and it is our strong belief that this is the case, then one would expect a great savings in the performance of such state-similarity algorithms.

Another natural extension is to apply this work to partially observable MDPs (POMDPs). A POMDP basically consists of an MDP in which the actual state of the system is hidden; instead one has a visible set of observations and a probabilistic observation function. Each POMDP induces a continuous MDP from which a solution may be recovered. In this sense, our results for continuous MDPs would immediately apply; however, a more direct solution would be preferable.

The most evident use of our metrics is in analyzing state aggregations; however, the original motivation for a quantitative notion of bisimulation was to study performance properties of a system, specified in terms of a modal logic (Desharnais et al., 1999; Desharnais, 2000). In fact, the original LMP metrics were defined in terms of a real-valued modal logic that captured properties of the system's states. Though we have not covered the logical approach for the continuous case in this work, it should easily be carried over with only slight modification. Thus, our metrics have a potential use in reasoning about logical properties of continuous MDPs too.

There has also been some preliminary work on knowledge transfer of policies in MDPs (Phillips, 2006). The basic idea is that if two MDPs have small overall bisimulation distance then Theorem 3.4.10 tells us that their optimal value functions, and hence optimal polices, should not be too far apart. One could potentially solve a class of MDPs by using the solution to a base MDP to which they are all similar, and modifying that policy accordingly.

Lastly, one might say the next logical step is to allow other parameters, e.g. time or the space of actions, to be continuous.

Bibliography

Alon, N., Ben-David, S., Cesa-Bianchi, N., & Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, *44*, 615–631.

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

Bernardo, M., & Bravetti, M. (2003). Performance measure sensitive congruences for Markovian process algebras. *Theoretical Computer Science*, *290*, 117–160.

Billingsley, P. (1968). *Convergence of probability measures.* Wiley.

Blute, R., Desharnais, J., Edalat, A., & Panangaden, P. (1997). Bisimulation for labelled markov processes. *LICS '97: Proceedings of the 12th Annual IEEE Symposium on Logic in Computer Science* (pp. 149–159). Washington, DC, USA: IEEE Computer Society.

Bouchard-Côté, A., Ferns, N., Panangaden, P., & Precup, D. (2005). An approximation algorithm for labelled markov processes: Towards realistic approximation. *QEST '05: Proceedings of the Second International Conference on the Quantitative Evaluation of Systems (QEST'05) on The Quantitative Evaluation of Systems* (pp. 54–61). Washington, DC, USA: IEEE Computer Society.

Boutilier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, *11*, 1–94.

Boutilier, C., Dearden, R., & Goldszmidt, M. (1995). Exploiting structure in policy construction. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1104–1111). San Francisco: Morgan Kaufmann.

Boutilier, C., Dearden, R., & Goldszmidt, M. (2000). Stochastic dynamic programming with factored representations. *Artificial Intelligence*, *121*, 49–107.

Clarkson, K. L. (2006). Nearest-neighbor searching and metric space dimensions. In G. Shakhnarovich, T. Darrell and P. Indyk (Eds.), *Nearest-neighbor methods for learning and vision: Theory and practice*, 15–59. MIT Press.

Dean, T., & Givan, R. (1997). Model minimization in markov decision processes. *Association for the Advancement of Artificial Intelligence AAAI/ Innovative Applications of Artificial Intelligence IAAI* (pp. 106–111).

Dean, T., Givan, R., & Leach, S. (1997). Model reduction techniques for computing approximately optimal solutions for markov decision processes. *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)* (pp. 124–131). San Francisco, CA: Morgan Kaufmann.

Dedecker, J., Prieur, C., & Raynaud De Fitte, P. (2004). Parametrized kantorovich-rubinstein theorem and application to the coupling of random variables. *ArXiv Mathematics e-prints.*

Desharnais, J. (2000). *Labelled markov processes.* Doctoral dissertation, McGill University. Adviser-Prakash Panangaden.

Desharnais, J., Gupta, V., Jagadeesan, R., & Panangaden, P. (1999). Metrics for labeled markov systems. *CONCUR '99: Proceedings of the 10th International Conference on Concurrency Theory* (pp. 258–273). London, UK: Springer-Verlag.

Desharnais, J., Gupta, V., Jagadeesan, R., & Panangaden, P. (2004). Metrics for labelled markov processes. *Theor. Comput. Sci.*, *318*, 323–354.

Desharnais, J., Jagadeesan, R., Gupta, V., & Panangaden, P. (2002). The metric analogue of weak bisimulation for probabilistic processes. *LICS '02: Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science, Copenhagen, Denmark, 22-25 July 2002* (pp. 413–422). Washington, DC, USA: IEEE Computer Society.

Dudley, R. M. (2002). *Real analysis and probability.* Cambridge University Press.

Dudley, R. M., Giné;, E., & Zinn, J. (1991). Uniform and universal glivenko-cantelli classes. *Journal of Theoretical Probability*, *4*, 485–510.

Edalat, A. (1997). When scott is weak on the top. *Mathematical Structures in Computer Science*, *7*, 401–417.

Even-Dar, E., & Mansour, Y. (2003). Approximate equivalence of markov decision processes. *16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings* (pp. 581–594). Springer.

Ferns, N. (2003). Metrics for markov decision processes. Master's thesis, McGill University, Montreal, Canada.

Ferns, N., Castro, P. S., Precup, D., & Panangaden, P. (2006). Methods for computing state similarity in markov decision processes. *Proceedings of the 22nd*

Annual Conference on Uncertainty in Artificial Intelligence (UAI-06). Arlington, Virginia: AUAI Press.

Ferns, N., Panangaden, P., & Precup, D. (2004). Metrics for finite markov decision processes. *AUAI '04: Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence* (pp. 162–169). Arlington, Virginia, United States: AUAI Press.

Ferns, N., Panangaden, P., & Precup, D. (2005). Metrics for markov decision processes with infinite state spaces. *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)* (pp. 201–208). Arlington, Virginia: AUAI Press.

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

Frangioni, A., & Manca, A. (2006). A computational study of cost reoptimization for min-cost flow problems. *INFORMS J. on Computing*, *18*, 61–70.

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

Givan, R., Dean, T., & Greig, M. (2003). Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, *147*, 163–223.

Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, *38*, 293–306.

Goto, J. H., Lewis, M. E., & Puterman, M. L. (2004). Coffee, tea, or ...?: A markov decision process model for airline meal provisioning. *Transportation Science*, *38*, 107–118.

Hennessy, M., & Milner, R. (1985). Algebraic laws for nondeterminism and concurrency. *Journal of the Association for Computing Machinery (JACM)*, *32*, 137–161.

Kozen, D. (1983). A probabilistic pdl. *STOC '83: Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing* (pp. 291–297). New York, NY, USA: ACM.

Lane, T., & Pack Kaelbling, L. (2002). Approaches to macro decompositions of large markov decision process planning problems. *Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference on Mobile Robots XVI* (pp. 104–113). Newton, MA.

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. *AI & MATH '06: Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics* (pp. 531–539). Fort Lauderdale, Florida, USA.

Likhachev, M., Gordon, G., & Thrun, S. (2005). Planning for markov decision processes with sparse stochasticity. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems 17*, 785–792. Cambridge, MA: MIT Press.

Milner, R. (1980). *A calculus of communicating systems*, vol. 92 of *Lecture Notes in Computer Science*. New York, NY: Springer-Verlag.

Milner, R. (1989). *Communication and concurrency*. Prentice-Hall International.

Müller, A. (1997). How does the value function of a markov decision process depend on the transition probabilities? *Mathematics of Operations Research*, *22*, 872–885.

Nedas, K. A. (2005). `http://www.spatial.maine.edu/~knedas/dev/soft/munkres.htm`.

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

Ortner, R. (2007). Pseudometrics for state aggregation in average reward markov decision processes. *Algorithmic Learning Theory: 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007, Proceedings Series: Lecture Notes in Computer Science , Vol. 4754 Sublibrary: Lecture Notes in Artificial Intelligence* (pp. 373–387). Springer-Verlag.

Park, D. (1981). Concurrency and automata on infinite sequences. *Proceedings of the 5th GI-Conference on Theoretical Computer Science* (pp. 167–183). London, UK: Springer-Verlag.

Parthasarathy, K. R. (1967). *Probability measures on metric spaces.* New York: Academic.

Phillips, C. (2006). Knowledge transfer in markov decision processes. `http://www.cra.org/Activities/craw/cdmp/awards/2006/Phillips/summary.pdf`. Final report for Canadian Distributed Mentors Project scholarship.

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

Rachev, S., & Rueschendorf, L. (1998). *Mass transportation problems. vol. 1: Theory. vol. 2: Applications.* Springer Series in Statistics. Probability and its Applications. New York: Springer.

Rudin, W. (1976). *Principles of mathematical analysis.* New York: McGraw-Hill. third edition.

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems* (pp. 1038–1044). The MIT Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

van Breugel, F., Sharma, B., & Worrell, J. (2007). Approximating a behavioural pseudometric without discount for probabilistic systems. *Foundations of Software Science and Computational Structures, 10th International Conference, FOSSACS 2007, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2007, Braga, Portugal, March 24-April 1, 2007, Proceedings* (pp. 123–137). Springer.

van Breugel, F., & Worrell, J. (2001a). Towards quantitative verification of probabilistic transition systems. *ICALP '01: Proceedings of the 28th International Colloquium on Automata, Languages and Programming,* (pp. 421–432). London, UK: Springer-Verlag.

van Breugel, F., & Worrell, J. (2001b). An algorithm for quantitative verification of probabilistic transition systems. *CONCUR '01: Proceedings of the 12th International Conference on Concurrency Theory* (pp. 336–350). London, UK:

Springer-Verlag.

Villani, C. (2002). Topics in mass transportation. `http://www.math.toronto.edu/hmaroofi/seminar/articles/Vilnotes.ps` (28/07/03). Preprint.

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

Webster, J. (2006). Finite approximation of measure and integration. *Annals of Pure and Applied Logic*, *137*, 439–449.

Winskel, G. (1993). *The formal semantics of programming languages.* Foundations of Computer Science Series. Cambridge, Mass.: MITP.