# Metrics for Markov Decision Processes

Norman Francis Ferns[1]

School of Computer Science

McGill University, Montreal

A thesis submitted to McGill University
in partial fulfilment of the requirements of
the degree of Master of Science
December 2003

[1] Research supported in part by FCAR

*Your file Votre référence*
*ISBN: 0-612-98632-2*
*Our file Notre référence*
*ISBN: 0-612-98632-2*

# Abstract

We present a class of metrics, defined on the state space of a finite Markov decision process (MDP), each of which is sound with respect to stochastic bisimulation, a notion of MDP state equivalence derived from the theory of concurrent processes. Such metrics are based on similar metrics developed in the context of labelled Markov processes, and like those, are suitable for state space aggregation. Furthermore, we restrict our attention to a subset of this class that is appropriate for certain reinforcement learning (RL) tasks, specifically, infinite horizon tasks with an expected total discounted reward optimality criterion. Given such an RL metric, we provide bounds relating it to the optimal value function of the original MDP as well as to the value function of the aggregate MDP. Finally, we present an algorithm for calculating such a metric up to a prescribed degree of accuracy and some empirical results.

# Résumé

Nous présentons une classe de métriques, définies sur l'espace des état d'un processus fini de décision de Markov (PFDM), chacune étant compatible avec la bisimulation stochastique; une notion d'équivalence d'état de PFDM dérivée de la théorie des processus parallèles. Ces métriques sont basées sur des métriques similaires développées dans le contexte des processus marqués de Markov. Comme ces derniers, les métriques conviennent à l'agrégation des états. De plus, nous limitons notre attention à un sous-ensemble de cette classe appropriée pour certaines tâches d'apprentissage par renforcement (APR); spécifiquement, les tâches à horizon infini, avec pour critère d'optimalité l'espérance de la récompense totale escomptée. Étant donné une telle métrique PFDM, nous donnons des bornes la comparant à la valeur optimale du PFDM original en plus de la fonction de valeur du PFDM agrégé. Finalement, nous présentons un algorithme pour calculer une telle métrique à n'importe quel degré de précision désiré ainsi que certains résultats expérimentaux.

# Acknowledgements

I owe an enormous debt of gratitude to each of the following people (and probably to many others whose names escape me at the moment) for helping me with this seemingly neverending thesis.

I'd like to thank my supervisor, Prakash Panangaden, for always providing significant insight into any problem with which I confront him, for always keeping me on my toes, and for being extremely patient with me. Rock on Prakmaster P!

I'd also like to thank my co-supervisor, Doina Precup, for always being able to provide a unique perspective to every problem we encounter, enabling me to understand concepts in a way I just never could on my own, for also being very patient with me, and for being probably the nicest person I have ever met. Thanks!

Many thanks are also due to the people in the School of Computer Science (particularly, the lovely ladies in the CS office) for providing a wonderful learning atmosphere and helping me wade through administrative matters.

I would like to thank the many friends I made here at McGill for helping me stay afloat. In particular, I would like to thank Marc Berndl, Bogdan (Boggy) Nica, Igor Khavkine, and Sébastien Loisel for the many helpful discussions.

A very special shout out goes to Spidey, May, Gobby, HCP, Doctor Zaius, Matthias and the Family, Villanos IV and V, Ruggero Deodato, Cthulhu in His house at R'lyeh, and Red-Faced Guy.

Last, but certainly not least, I owe many thanks to my family, including my brother Ryan Ferns, my mother Ruth Pinto Ferns, my sister Raquel

Ferns, Pebbles the bird, and my father Norman Ferns Sr. for, among other things, the love, the support, the inspiration, the food, the money, waking me up in the morning, the food again (it really is that good Mamma), and for not kicking me out of the house.

This work is dedicated to all the oddball-slacker-loser-weirdos out there.

# Contents

# Chapter 1

# Introduction

Consider each of the following scenarios:

- *Mobile Robotics[11]*

  An autonomous mobile robotic agent must perform the following task: deliver, in a timely manner, a fixed number of packages of varying priorities to different locations on the same floor of a certain building, while attempting to respect the relative package priorities. The robot has access to a 2D map of the location, which is assumed to be discretized into a grid. Within each grid cell, the robot may choose to move in one of four directions (NORTH, SOUTH, EAST, WEST) with the provision that actions leading the robot off the grid leave the robot's position unchanged. However, there is a small probability of error, i.e. if the robot chooses to move into a certain grid, it may actually move into another grid adjacent to its current position or even not move at all.

- *Airline Meal Provisioning[9]*

  Airline meal provisioning involves the production of meals on the ground prior to flight departure for eventual inflight passenger service. An important question is to determine the quantity of meals that should be prepared beforehand in order to achieve high passenger satisfaction while maintaining low costs. More specifically, we have the following simplified decision-making scenario: long before flight departure a fixed number of meals is prepared. At several subsequent decision epochs prior to flight departure, the meal quantity may be adjusted, i.e. meals may be added or subtracted. However, meals added closer to the time of departure are more costly. Here, one aims to find a policy of optimal adjustment for each of these decision epochs.

- *Game of Tetris[25]*

  In the game of tetris, "bricks" fall one at a time into a rectangular grid. The horizontal position and orientation of each piece is chosen by the user as it falls, until it hits the current pile of "bricks" or the bottom of the grid. If the resultant configuration fills an entire row of the grid, the row of "bricks" in that row is eliminated and the next piece falls. The game continues in this manner, unless the pile of pieces reaches the top of the grid, in which case the game is over. The goal of the game is to continually choose the position and orientation of pieces so as to eliminate the most rows (or equivalently, to acquire the most game points).

Underlying each of these situations is the notion of a Markov decision process (see [20]). A Markov decision process (MDP) is a model for discrete-time sequential decision making under uncertainty used widely throughout such fields as operations research, control theory, electrical engineering, and computer science. It is characterized by a set of states, a set of actions available in each state, (Markovian) state transition probabilities for each action, and numerical rewards or costs associated with each state and action.

For example, one can model a game of tetris by taking as a state a particular configuration of a pile of "bricks" and a falling piece, as an action available in that state an orientation and a horizontal position for the falling piece, and as a reward the number of rows eliminated by choosing that action in that state. Transition probabilities can be set by determining the impact of the falling piece and noting that the next falling piece is generated at random.

Given an MDP, one seeks to solve it by finding a strategy that dictates the "best" action to take in each state. Of course, what is meant by "best" varies with each problem. In tetris, for example, one seeks a strategy for choosing piece positions and orientations in order to eliminate the most rows in the long run. In reinforcement learning (RL) one generally seeks a strategy that maximizes longterm reward (or minimizes longterm cost). Fortunately, there exist many RL algorithms to solve these types of problems in time polynomial in the size of the input. Unfortunately, it is often the case that these methods become impractical for real-world problems with large state space models. It is therefore highly desirable to model problems with MDPs of small size.

In this work we introduce a family of distance measures, or metrics, on the states of an MDP with finite state and action spaces with the following property: each metric assigns to a pair of states a distance that is inversely proportional to how "similar" those states are. Such metrics could potentially be used for lossy MDP state compression by clustering states which are within a small distance of each other. Of course, all of these concepts need to be formalized.

Before proceeding let us note that it is our ultimate goal to extend such tools to handle continuous state space MDPs; however, as is usually the case, it is worthwhile to begin our investigation in a suitably simplified setting, that being the realm of finite state space models.

## 1.1 Outline

The result of this work is outlined as follows:

- In chapter 2 we formally introduce the basics of finite Markov decision processes, reinforcement learning, and metrics, as well as recent work in metric based compression of certain probabilistic systems.
- In chapter 3 we derive a family of metrics for finite MDPs and provide a polynomial time algorithm to compute a subset of such metrics useful for certain reinforcement learning tasks.
- In chapter 4 we provide experimental results demonstrating the relationship between the metrics and solutions to various finite MDPs.
- In chapter 5 we conclude by discussing related and future work.

# Chapter 2

# Background

In this chapter we will provide background on finite Markov decision processes in the context of reinforcement learning, as well as metric-based compression of a related model. The reader is assumed to have a basic knowledge of probability and measure theory.

## 2.1 Markov Decision Processes

Consider the sequential decision model represented in figure 2.1 (originally from section 3 of [1]), depicting the interaction between a decision-maker, or agent, and its environment. We assume that time is discrete, not necessarily consisting of equal units but rather a sequence of discrete decision epochs. 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

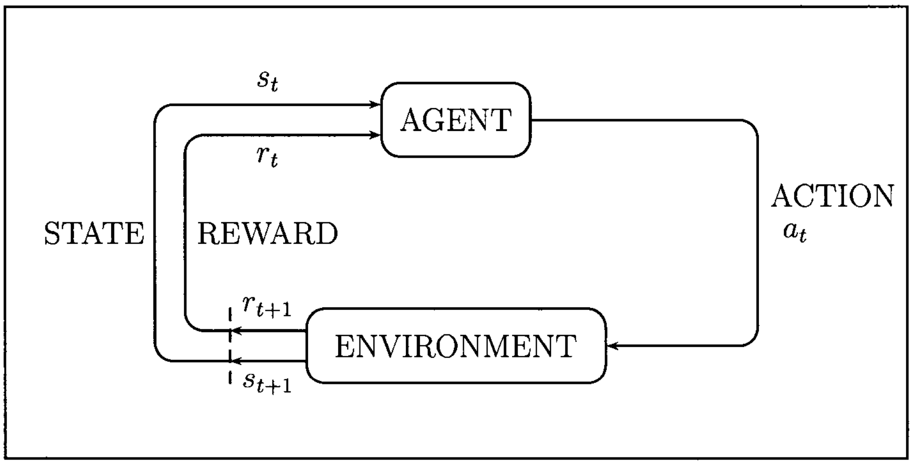

Figure 2.1: Agent-Environment Interaction.

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 $r_{t+1}$ (interpreted as a reward or a cost) from the environment and the environment evolves to a new state according to a probability distribution induced by $a_t$. The agent then 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.

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* (MDP). We will restrict our attention to the case where both the state space and the action space are finite. While our goal, ultimately, is to apply metric-based compression to continuous state space MDPs, the finite state space MDP, lacking any measure theoretic hassles, is an appropriately simple place to begin our investigation. Moreover, compression of large finite state space MDPs is itself of great importance due to the dependence of the running times of algorithms used to "solve" MDPs and the sizes of those MDPs.

Formally, we have the following:

**Definition 2.1.1.** 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 $s$
- $\forall s \in S, \forall a \in A_s, P(\cdot|s,a) : S \to [0,1]$ is a stationary Markovian subprobability transition function. $\forall s' \in S, P(s'|s,a)$ is the probability of transitioning from state $s$ to state $s'$ under action $a$ and will be denoted by $P^a_{ss'}$.
- $\forall s \in S, \forall a \in A_s, r(s,a)$ is the immediate reward associated with choosing action $a$ in state $s$.

Allowing $P(\cdot|s,a)$ to be a subprobability function, i.e $P(S|s,a)$ may have total mass less than 1, means that there may be a positive probability of not transitioning, given of course by $1 - P(S|s,a)$. We will make the common assumption that this is not the case, i.e. for every $s$ and $a$, $P(\cdot|s,a)$ is a full probability function.

A finite MDP (hereafter, MDP) can also be specified via a state-transition diagram as seen in figure 2.2.

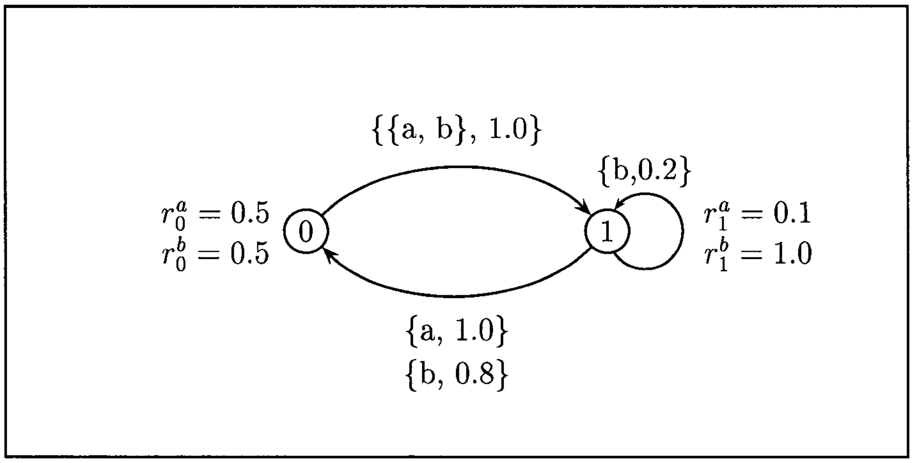

Figure 2.2: A finite state MDP with 2 states labeled 0 and 1, and 2 actions labeled $a$ and $b$. On action $a$ state 0 yields a reward of 0.5, state 1 yields a reward of 0.1, and each transitions with probability 1.0 to the other. On action $b$ state 0 once more yields a reward of 0.5 and transitions to state 1 with probability 1.0. However, state 1 yields a reward of 1.0, transitions to state 0 with probability 0.8, and remains in state 1 with probability 0.2.

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 *Reinforcement Learning Problem.*

## 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, also known as neuro-dynamic programming, to be that branch of AI that deals with an agent learning through interaction with its environment in order to achieve a goal. This interaction is exactly as described in the Markov decision process framework above. Here we think of the numerical signal received by the agent as a means of providing it with a reward or a punishment as a direct consequence of its actions, thereby enabling it to learn which action selection strategies are good and which are bad via its own behaviour. The optimality criterion of a reinforcement learning problem is aimed roughly towards maximizing the cumulative reward achieved throughout the course of interaction, i.e. favours strategies that are beneficial in the long run. 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.

### 2.2.1 Policies

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, we allow for policies that are stochastic, history-dependent, and even non-stationary. These characteristics lead to six classes of policies which we will formally introduce shortly. First we introduce some notation.

A *trajectory* is an alternating sequence of states and actions $s_0a_0s_1a_1\ldots$. A *history* is a trajectory of length $n \in \mathbb{N}$ denoted $h_n = s_0a_0s_1a_1\ldots s_{n-1}a_{n-1}s_n$. The space of all histories of length $n$ is $H_n = (S \times A)^n \times S$.

**Definition 2.2.1.** A *randomized policy* is a sequence of mappings $\{\pi_n\}$ where $\pi_n : H_n \times A \to [0,1]$ is such that $\pi_n(h_n, a)$ is the probability of choosing action $a$ given history $h_n$, and $\pi_n(h_n, A_{s_n}) = 1$.

**Definition 2.2.2.** A policy is said to be:

(a) *randomized Markov* if each mapping depends only on the current state, in which case $\pi_n : S \times A \to [0,1]$.

(b) *randomized stationary* if each mapping is time independent, in which case it consists of a single mapping $\pi : S \times A \to [0,1]$.

(c) *deterministic* if each history completely determines an action, in which case $\pi_n : H_n \to A$ and $\pi_n(h_n) \in A_{s_n}$.

(d) *deterministic Markov* if it is deterministic and each mapping depends only on the current state, in which case $\pi_n : S \to A$.

(e) *deterministic stationary* if it is deterministic and each mapping is time independent, in which case it consists of a single mapping $\pi : S \to A$.

We denote the class of all randomized, randomized Markov, randomized stationary, deterministic, deterministic Markov, and deterministic stationary policies by $\Pi^R$, $\Pi^{RM}$, $\Pi^{RS}$, $\Pi$, $\Pi^M$, and $\Pi^S$ respectively. Clearly we have

$$\Pi^S \subseteq \Pi^M \subseteq \Pi$$

$$\Pi^{RS} \subseteq \Pi^{RM} \subseteq \Pi^R$$

$$\Pi \subseteq \Pi^R, \Pi^M \subseteq \Pi^{RM}, \Pi^S \subseteq \Pi^{RS}$$

We will also denote an arbitrary policy by $\pi$; its explicit form (as a sequence of mappings or as a single mapping) will be clear from the context.

Suppose we are given an initial distribution on states $P_0$ (for example, one might take $P_0$ to be the Dirac measure at some initial state $s$, i.e $\delta_s$ where $\delta_s(A)$ is 1 if $A$ contains $s$ and 0 otherwise). Then any policy $\pi$ induces a distribution on the space of histories via

$$P^\pi(h_n) = P_0(s_0) \prod_{k=0}^{n-1} \pi_k(h_k, a_k) P^{a_k}_{s_k s_{k+1}}$$

(which may be simplified in case $\pi$ is Markov, stationary, or deterministic). These probabilities may be extended to conditional probabilities (independent of $P_0$) in the usual way (for a formal development the reader is directed to section 2.1.6 of [20]) and so it makes sense to speak of the conditional expectation given $\pi$. We will use this to formalize the optimality criteria for RL tasks.

### 2.2.2 Optimality Criteria

We have previously stated that the optimality criterion of the RL problem is concerned with maximizing the sum of the sequence of numerical rewards obtained via the agent's interaction with the environment. More specifically, a RL task involves finding a policy $\pi$ that maximizes for every state $s \in S$:

1. *finite horizon (episodic task)*

$$E^{\pi}[R_t | s_t = s]$$

$$\text{where the } \textit{return } R_t = \sum_{k=0}^{T-(t+1)} r_{t+k+1}$$

2. *infinite horizon (continual task)*

   (a) *total reward*

$$\lim_{T \to \infty} E^{\pi}[R_t | s_t = s]$$

   (b) *average reward*

$$\lim_{T \to \infty} \frac{1}{T} E^{\pi}[R_t | s_t = s]$$

   (c) *total discounted reward*

$$\lim_{T \to \infty} E^{\pi}[R_t | s_t = s]$$

$$\text{where } R_t = \sum_{k=0}^{T-(t+1)} \gamma^k r_{t+k+1} \text{ for some } \gamma \in [0, 1).$$

whenever each exists and is finite. In these cases such a maximizing policy is said to be *optimal* and in the infinite horizon case *total reward optimal*, *average reward* or *gain optimal*, or *discount optimal*, depending on the particular criterion used.

For the most part, we will focus on infinite horizon tasks, as is the case in much of the research in the RL community. Our reasons for doing so are many. For example, if the horizon, $T$, is not too big then there exists an efficient solution to the problem of determining an optimal policy (as discussed in section 2.2.4). More troublesome is the fact that optimal policies for an episodic task are generally nonstationary (see [20]), which is somewhat impractical in real world planning or control operations, where an easy-to-implement policy is of great importance.

As far as infinite horizon problems are concerned, the total reward criterion, while formalizing exactly what we desire, is problematic since there is no guarantee that the limit exits. For example, even in a finite MDP, where rewards are necessarily bounded, it is possible for the return we are attempting to maximize to be infinite. To get around this problem, researchers have turned to the average and discounted reward models.

With an average reward criterion one seeks to maximize the long term average cumulative reward. This criterion is unattractive for many reasons. The limit may fail to exist even in the case of a finite MDP (see section 8.1.1 of [20]). When limits do exist, average reward optimality may still fail to distinguish between gain optimal policies that are behaviourally quite different. As a result a more advanced analysis is required (one looks for biased-gain optimal policies), which (while not technically difficult) is quite laborious. Moreover, the average reward model can be expressed as the limit of the expected total discounted reward model as the *discount factor* $\gamma$ tends to 1. The latter model is technically pleasing and quite easy to handle; so, while there exist situations in which an average reward criterion is the model of choice, for the most part one works with the discounted model.

The total discounted reward criterion involves geometrically discounting the sequence of rewards obtained. The reasoning behind discounting 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 mathematical tractability and existence of stationary optimal policies (see [20]), resulting in its being the dominant model used for RL tasks. For these reasons, we will mainly concentrate on the discounted model in the rest of this work.

We conclude with a simple example (from [12]) demonstrating the differences between the optimality criteria just discussed. Consider the finite MDP in figure 2.2.2 with 14 states and 3 actions. All transitions have probability 1 for the specified actions and the unlabeled actions are assumed to be $a$. [1] All rewards are assumed to be zero, except as indicated in the diagram. It is evident that the only real choice occurs in state $s$, where actions $a$, $b$, and $c$ each restrict the agent to one of the three chains. Consider a finite-horizon model in which decisions are made for the first $H$ steps only. With $H = 5$ the three actions $a$, $b$, and $c$ yield returns of 6, 0, and 0 respectively, so that action $a$ is optimal. Under a discounted criterion with $\gamma = 0.2$ the same three actions yield returns of 0.1, 0.004, and 0.00088, so that action $a$ is still optimal. However, if we choose $H = 15$ in the finite-horizon model then the three returns are 26, 100, and 99, making action $b$ optimal. If we take $\gamma = 0.9$ in the discounted model then the three returns are 16.2, 59.049, and 58.45851, again making action $b$ optimal. On the other hand, if we use an

[1] To ensure that all induced distributions have total mass 1 we may add an absorbing state which transitions to itself on all actions and receives a reward of zero, and then add transitions to this state as needed. Clearly, this does not affect the optimal actions for $s$ under the different criteria.

average reward criterion the expected average returns are 2, 10, and 11, so that action $c$ is optimal.

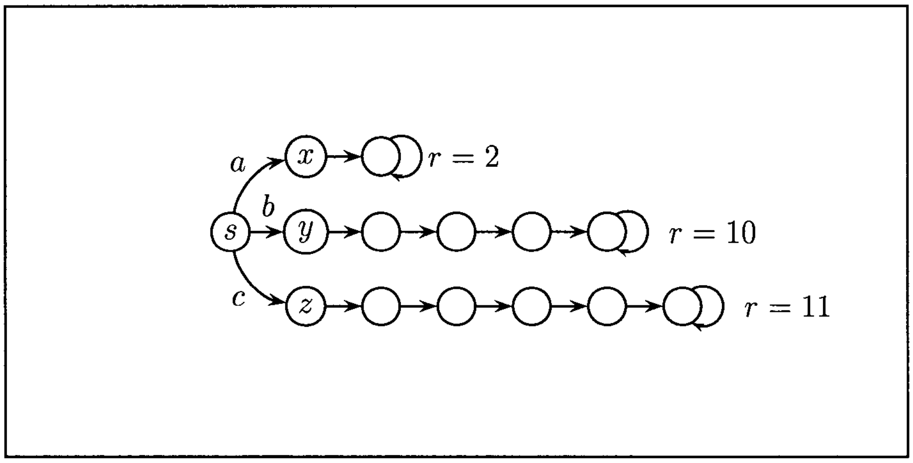

Figure 2.3: An MDP with different optimality criteria. In state $s$, action $a$ is optimal if $H = 5$ but action $b$ is optimal if $H = 15$ in the finite horizon model. In the discounted reward model, action $b$ is optimal if $\gamma = 0.9$ but action $a$ is optimal if $\gamma = 0.2$. In the average reward model, action $c$ is optimal.

### 2.2.3 The Value of a Policy

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

In terms of value functions, a policy $\pi^*$ is optimal iff $V^{\pi^*}(s) \geq V^{\pi}(s)$ for every $s \in S$ and $\pi \in \Pi^R$. As we have mentioned above, a valuable fact about infinite horizon discounted models is that an optimal policy always exists. Moreover, a stationary optimal policy always exists. Thus, we need only examine the space of stationary policies in our search for an optimal policy. Actually, we will restrict our attention to the space of randomized stationary policies, $\Pi^{RS}$.

Given $\pi \in \Pi^{RS}$, we can use the Markov property to derive for any $s \in S$:

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

The fixed point 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 $\pi$, and it is a theorem that $V^\pi$ 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 $\pi^*$, 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 also the unique solution (see sections 6.1 and 6.2 of [20]). These are known as the *Bellman optimality equations.*

### 2.2.4 Policy Evaluation and Value 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 be used as a backup rule for value function computation. This is a essentially a consequence of the Banach fixed point theorem, and is known respectively as policy evaluation and value iteration:

**Theorem 2.2.3.** *Let* $\pi \in \Pi^{RS}$. *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* $V^\pi$ *uniformly.*

**Theorem 2.2.4.** *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 can be realized via a dynamic programming (DP) algorithm that computes value functions up to a prescribed degree of accuracy. For example, if one is given a positive error $\epsilon$ then iterating until the maximum difference between consecutive iterates is $\frac{\epsilon(1-\lambda)}{2\lambda}$ guarantees that the current iterate differs from the true value function by at most $\epsilon$ (for details see section 6.3.2 of [20]).

The DP algorithm for value iteration is a standard RL solution method; many alternative solution methods are based on it while aiming to improve computational efficiency. The problem with the DP algorithm is that, while it is polynomial in $|S|$ and $|A|$, it is also subject to the *curse of dimensionality*: roughly, if we think of the state space as being generated by state variables then the computational costs involved increase exponentially with an increase in state variables. In general, we should still expect such methods to be impractical when dealing with fairly large state spaces, as the entire state space must be examined.

It would, therefore, be of great importance to be able to express a particular model via a behaviourally equivalent model of significantly reduced size, so that one could obtain information about the original model by solving the reduced model. Various methods have been proposed as to how this should best be done; in the next few sections we will see how distances can be assigned to states of a particular finite state system, allowing for the use of distance-based compression.

## 2.3 Labeled Markov Processes

As previously stated, our goal is to efficiently and accurately compress the state space of an MDP. Then an RL problem can be solved using the compressed MDP. We will utilize a method of state space reduction first employed in the context of labeled Markov processes (LMPs) [2].

**Definition 2.3.1.** A *labeled Markov process* is a quadruple

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

where:

- $S$ is an *analytic* [2] set of states
- $\Sigma$ is the Borel $\sigma$-field on $S$
- $A$ is a finite set of actions
- $\forall a \in A, \tau_a : S \times \Sigma \to [0, 1]$ is a stationary subprobability transition kernel, i.e. $\forall X \in \Sigma, \tau_a(\cdot, X)$ is a measurable function and $\forall s \in S, \tau_a(s, \cdot)$ is a subprobability measure

LMPs are roughly (continuous state space) MDPs without rewards (note that an MDP can equivalently be formulated simply by specifying the entire action space $A$, rather than the action set $A_s$ for each state $s \in S$).

[2] An analytic set is the continuous image of a Polish space under a map between Polish spaces. A Polish space is a topological space homeomorphic to a complete separable metric space. For more information see section 13.2 of [6].

### 2.3.1 Bisimulation for LMPs

The notion of LMP state equivalence used in [4] originates from the theory of concurrent processes, and is known as *bisimulation* [18]. Milner utilized a strong bisimulation in [16] 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* [10]; two processes are bisimilar iff they satisfy precisely the same formulas.

In [15], Larsen and Skou extended this notion to a probabilistic framework. Their *probabilistic bisimulation* was developed as an equivalence notion for labeled Markov chains (LMCs), the discrete version of LMPs. They provide characterizations of probabilistic bisimulation both in terms of a maximal matching relation and a probabilistic modal logic.

For LMPs we consider strong probabilistic bisimulation. Given a relation $R$ on $S$, a subset $X$ of $S$ is said to be $R$-closed iff $\{s' \in S | \exists s \in X.\ sRs'\} \subseteq X$.

**Definition 2.3.2.** A *(strong probabilistic) bisimulation relation* is an equivalence relation on $S$ that satisfies the following property:

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

We say states two states are *(strongly probabilistic) bisimilar* iff they are related by some bisimulation relation.

Bisimulation is essentially the largest of the bisimulation relations, where each such relation relates those states that can be "behaviourally lumped" together. In other words, an outside observer witnessing the behaviours of an LMP and the reduced state space LMP in which all bisimilar states have been combined, could not distinguish between the two systems.

Unfortunately, bisimulation is too restrictive. Consider as an example the LMC in figure 2.4 taken from [24]. The states 0 and 1 are bisimilar iff $\epsilon = 0$, since each must have the same probability of transitioning to the class containing state 4. However, if $\epsilon$ is very close to 0 then we should expect the systems to be almost bisimilar, i.e. they should mostly behave the same way. This is particularly striking when one considers that in most systems one works with approximations of the system parameters rather than with the exact values of the parameters themselves. A quantitative notion of bisimulation is needed. The appropriate notion is found in the realm of

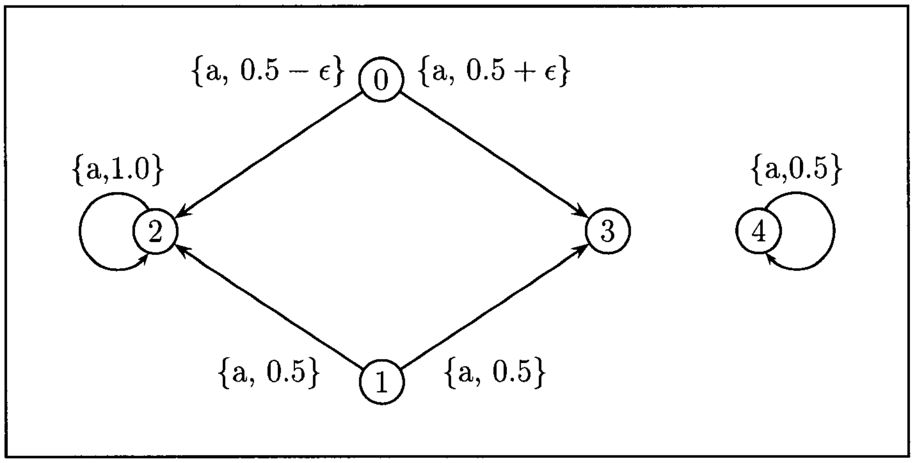

Figure 2.4: States 0 and 1 are bisimilar iff $\epsilon = 0$, where $\epsilon \in [-0.5, 0.5]$.

metrics.

## 2.4 Metrics

A metric is essentially a function that assigns distances between points in a space.

**Definition 2.4.1.** A *pseudometric* on a set $S$ is a map $d : S \times S \to [0, \infty)$ such that for all $s$, $s'$, $s''$:

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

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

Note that every pseudometric $d$ induces an equivalence relation on the set $S$, obtained by equating points assigned distance zero by $d$. Applying $d$ in the obvious way to the quotient space under this relation yields a full-fledged metric.

For convenience in the rest of this work we will use the words “metric” and “pseudometric” interchangeably; however, should the need arise to use “metric” in its proper sense as defined above, we will point this out.

**Definition 2.4.2.** Let $\mathcal{M}$ be the class of 1-bounded pseudometrics on $S$. We define a partial ordering on $\mathcal{M}$ as follows:

$$m_1 \preceq m_2 \Leftrightarrow \forall s, s' \in S, m_1(s, s') \geq m_2(s, s')$$

**Lemma 2.4.3.** $(\mathcal{M}, \preceq)$ *is a complete lattice.*

*Proof.* The least element, $\bot$, is given by

$$\bot(s, s') = \begin{cases} 0 & \text{if } s = s' \\ 1 & \text{otherwise} \end{cases}$$

and the greatest, $\top$, by $\forall s, s', \top(s, s') = 0$. Greatest lower bounds are given by $(\sqcap\{m_i\})(s, s') = \sup\{m_i(s, s')\}$. □

An important part of calculating distances between states for the models in consideration involves calculating the distances between the induced state distributions. In order to do so, we need some mechanism for moving metrics on a state space to metrics on the set of distributions on the state space.

### 2.4.1 Metrics Applied to Distributions

There are many ways to extend metrics on a space $S$ to metrics on the space of probability measures on $S$ (see, for example, [7]). The particular metric we will use is known variously as the Monge-Kantorovich distance, the Kantorovich-Rubinstein distance, the Hutchinson distance, and the bounded Wasserstein distance. We will refer to it simply as the *Kantorovich metric.*

The Kantorovich metric arose in the study of the *Monge-Kantorovich optimal mass transportation problem*(see [26]): Assume we are given a pile of sand and a hole, occupying measurable spaces $(X, \Sigma_X)$ and $(Y, \Sigma_Y)$ (see figure 2.5). The pile of sand and the hole are assumed to have the same volume, and the mass of the pile is assumed to be normalized to 1. Let $\mu$ and $\nu$ be measures on $X$ and $Y$ respectively, such that whenever $A \in \Sigma_X$

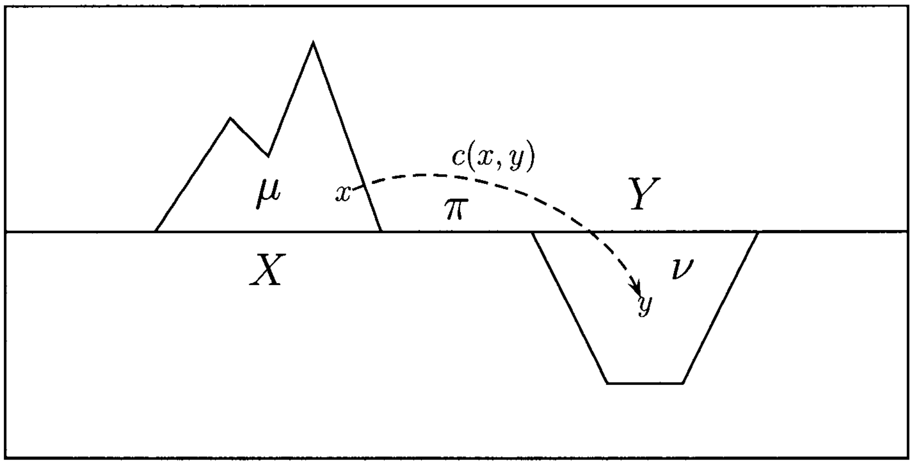

Figure 2.5: Kantorovich Optimal Mass Transportation Problem.

and $B \in \Sigma_Y$, $\mu[A]$ measures how much sand occupies $A$ and $\nu[B]$ measures how much sand can be piled into $B$. Suppose further that we have some measurable cost function $c : X \times Y \to \mathbb{R}$, where $c(x,y)$ tells us how much it costs to transfer one unit of mass from a point $x \in X$ to a point $y \in Y$. 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 $\pi$ on $(X \times Y, \Sigma_X \otimes \Sigma_Y)$, where $d\pi(x,y)$ measures how much mass is transferred from location $x$ to $y$. Of course, for the plan to be valid we require that $\pi[A \times Y] = \mu[A]$ and $\pi[X \times B] = \nu[B]$ for all measurable $A$ and $B$. A plan satisfying this condition is said to have marginals $\mu$ and $\nu$, and we denote the collection of all such plans by $\Pi(\mu,\nu)$. We can now restate the goal formally as:

$$\text{minimize} \int_{X \times Y} c \, d\pi \text{ over } \pi \in \Pi(\mu,\nu)$$

This is actually an instance of an infinite linear program. Fortunately, under very general circumstances, it has a solution and admits a dual formulation. This is true, for example, whenever at least one of $\mu$ and $\nu$ is *perfect*[3] and $c$ is a bounded measurable cost function (see [22]). The general form of duality from [22] states that for any bounded measurable cost function $c$,

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c d\pi = \sup_{(\phi,\psi) \in L'(c)} \int_X \phi d\mu + \int_Y \psi d\nu$$

where $L'(c)$ consists of all pairs of measurable functions $(\phi, \psi)$, integrable with respect to $\mu$ and $\nu$ respectively, such that for every $(x, y) \in X \times Y, \phi(x) + \psi(y) \leq c(x, y)$. If $S$ and $c$ satisfy certain regularity conditions, and $c$ is additionally a metric, then even more can be said. Assume, for example, that $X = Y = S$ is an analytic space equipped with its Borel $\sigma$-field $\Sigma$. Then every probability measure on $(S, \Sigma)$ is perfect.[4] Let $c$ be a bounded measurable pseudometric on $S$ and let $Lip(c)$ be the set of measurable Lipschitz functions mapping $S$ to the unit interval, i.e. $f : S \to [0, 1]$ with $f(s) - f(s') \leq c(s, s')$ for every $s, s' \in S$. Then from [22] and chapter 4 of [21] we have

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{S \times S} c d\pi = \sup_{f \in Lip(c)} \int_S f d(\mu - \nu)$$

and both extrema are attained. This motivates the following definition.

**Definition 2.4.4.** Given measurable $m \in \mathcal{M}$, the Kantorovich metric ap-

[3] A probability measure $P$ on $(S, \Sigma)$ is perfect iff for every real valued $\Sigma$-measurable function $f$ there exists a Borel set $B_f$ of $\mathbb{R}$ such that $B_f \subseteq f(S)$ and $P(f^{-1}(B_f)) = 1$. For more information see chapter II, section 4 of [19].

[4] Every probability measure on a Polish space is tight and therefore perfect. Moreover, every measure on an analytic space can be lifted to (i.e., is the the image measure of) a measure on a Polish space. It follows that every probability measure on an analytic space is perfect. For more information see theorem 4.1 of [19].

plied to $\mu$ and $\nu$ is defined as:

$$\sup_{f \in Lip(m)} (\int_S f d\mu - \int_S f d\nu)$$

and is denoted by $m(\mu, \nu)$.

The fact that this defines a 1-bounded pseudometric on the space of probability measures on $(S, \Sigma)$ is easy to see; for symmetry simply note that $f$ is non-expansive iff $1 - f$ is non-expansive. In addition, this definition may be extended to subprobability distributions by adding the requirement that $\mu(S) \geq \nu(S)$; however, duality in this case does not necessarily take the form described above.

What happens when we try to compute the Kantorovich metric using an affine transformation of a metric $m$, i.e. with a cost $c_1 m(x, y) + c_2$ for some constants $c_1$, $c_2$, with $c_1$ nonnegative? In this case the general form of duality yields:

**Lemma 2.4.5.**

$$\sup_{f \in Lip(c_1 m + c_2)} (\int_S f d\mu - \int_S f d\nu) \leq c_1 m(\mu, \nu) + c_2$$

*Proof.* Here we use the general form of duality and simply note that $c_1 m(x, y) + c_2$ is a bounded measurable cost function and for every $f \in Lip(c_1 m + c_2)$, $(f, -f) \in L'(c_1 m + c_2)$. Finally,

$$
\begin{aligned}
\sup_{f \in Lip(c_1 m + c_2)} \int_S f d(\mu - \nu) &\leq \sup_{(\phi,\psi) \in L'(c_1 m + c_2)} \int_S \phi d\mu + \int_S \psi d\nu \\
&= \inf_{\pi \in \Pi(\mu,\nu)} \int_{S \times S} (c_1 m + c_2) d\pi \\
&= c_1 \inf_{\pi \in \Pi(\mu,\nu)} \int_{S \times S} m d\pi + c_2 \\
&= c_1 \sup_{f \in Lip(m)} \int_S f d(\mu - \nu) + c_2 \\
&= c_1 m(\mu, \nu) + c_2
\end{aligned}
$$

□

We now note that in the case of a finite state space, measurability conditions are no longer required. Here, noting that integration reduces to summation, we see that the Kantorovich metric applied to state distributions $P$ and $Q$ is given by the following linear program:

$$
\begin{aligned}
&\max_{u_i} \sum_{i=1}^{|S|} (P(s_i) - Q(s_i)) u_i \\
\text{subject to: } &\forall i, j, u_i - u_j \leq m(s_i, s_j) \\
&\forall i, 0 \leq u_i \leq 1
\end{aligned}
$$

Its dual is given by

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

The discrete minimization program has an interpretation as a *Hitchcock transportation problem.* This is a specialized minimum cost network flow problem as depicted in figure 2.6.

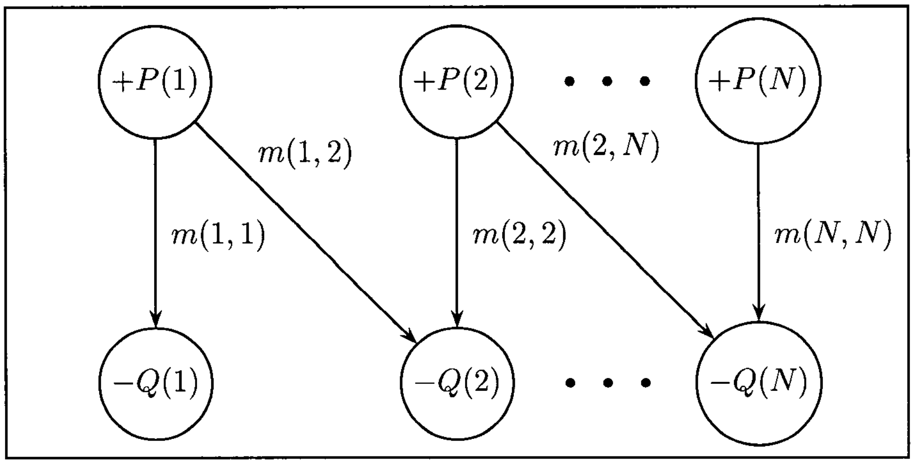

Figure 2.6: Hitchcock Network Transportation Problem ($N = |S|$).

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, say labelled $P(S)$ and $Q(s')$ for some $s$, $s' \in S$, respectively, there is a transportation arc labelled with the cost of transporting one unit from the source to sink, given here by $m(s, s')$. A flow is an assignment of the number (nonnegative) of units to be shipped along all arcs. We require 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. We also require 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.

An immediate consequence of this interpretation is that the Kantorovich metric in the discrete case is computable (assuming $m$ is computable), as linear programming techniques such as the network simplex method may be used to compute it. In fact, there exist strongly polynomial algorithms to solve it. For an up to date account of minimum cost flow algorithms we refer the reader to [27]. The fastest strongly polynomial algorithm is originally due to Orlin (see [17]) and has worst case running time $O(m \log m(m + n \log n))$ where $n$ is the number of nodes and $m$ is the number of arcs. Here, $n = 2|S|$ and $m = n^2$, so that the Kantorovich distance can be computed in time $O(|S|^2 \log |S|)$.

Thus, we can move metrics on the state space to metrics on distributions. Moreover, the ordering on metrics is preserved while doing so, as is easily seen from the dual program.

**Lemma 2.4.6.** *Suppose $m_1, m_2 \in \mathcal{M}$ with $m_1 \preceq m_2$. Then for all distributions $P$ and $Q$, $m_1(P,Q) \geq m_2(P,Q)$.*

The result above tells us that given a collection of 1-bounded metrics $\{m_i\}$ on $S$, we can move these to the complete lattice of 1-bounded metrics on the space of distributions on $S$, $\mathcal{P}(S)$. Here, as before, the greatest lower bound is given by $(\sqcap_{\mathcal{P}(S)}\{m_i\})(P,Q) = \sup\{m_i(P,Q)\}$. Naturally, the question arises as to whether this metric is induced by the corresponding metric in $\mathcal{M}$. In other words, is it true that $\sqcap_{\mathcal{P}(S)}\{m_i\}$ is the Kantorovich metric of $\sqcap_{\mathcal{M}}\{m_i\}$? Under certain conditions, we show that this is so.

**Lemma 2.4.7.** *Let $\{m_i\} \subseteq \mathcal{M}$ be a monotone decreasing sequence. Then for any distributions $P$ and $Q$, $(\sqcap\{m_i\})(P,Q) = \sqcup m_i(P,Q)$.*

*Proof.* It is obvious that $\sqcup m_i(P,Q) \leq (\sqcap\{m_i\})(P,Q)$, since a feasible solution for the primal LP for $m_i(P,Q)$ is a feasible solution for the primal LP for $(\sqcap\{m_i\})(P,Q)$ for every $i$.

For the other inequality we use the dual LP. For every $i$ let $l_{kj}^{(i)}$ be a feasible solution yielding the minimum for $m_i(P,Q)$. Note that for every $i$ each constitutes a feasible solution for the dual LP for $(\sqcap\{m_i\})(P,Q)$. Define $\epsilon_{kj}^{(i)} = (\sqcap\{m_i\})(s_k,s_j) - m_i(s_k,s_j)$ and $\delta_{kj} = \min(P(s_k), Q(s_j))$. Then for every $k$, $j$, and $i$:

- $\varepsilon_{kj}^{(i)} \geq 0$ and $\lim_{i\to\infty} \varepsilon_{kj}^{(i)} = 0$
- $l_{kj}^{(i)} \leq \delta_{kj}$.

Thus,

$$\begin{aligned}
(\sqcap\{m_i\})(P,Q) &\leq \sum_{k,j} l_{kj}^{(i)}(\sqcap\{m_i\})(s_k,s_j) \\
&= m_i(P,Q) + \sum_{k,j} l_{kj}^{(i)}\varepsilon_{kj}^{(i)} \\
&\leq \sqcup m_i(P,Q) + \sum_{k,j} \delta_{kj}\varepsilon_{kj}^{(i)}
\end{aligned}$$

By taking $i \to \infty$ on both sides of the inequality above we obtain

$$(\sqcap\{m_i\})(P,Q) \leq \sqcup m_i(P,Q)$$

as required. □

We are now ready to examine the role of metrics in the compression of LMPs.

## 2.5 Metrics for LMPs

In [4] the authors devised a collection of bisimulation metrics, metrics assigning distance zero to states iff they are bisimilar, on the state space of an LMP. They did so via a real valued modal logic characterizing bisimulation. On the other hand, in [23] the authors used methods from category theory to develop a fixed point characterization of these bisimulation metrics (in addition to others). We will present both characterizations in the context of LMCs; however, our presentation of the fixed point formulation will follow that given in [5], utilizing domain theoretical results in favour of categorical ones. In that work the system of interest is the labeled concurrent Markov chain (LCMC), roughly an LMC in which some states are allowed to make nondeterministic transitions.

For our presentation we fix an LMC $(S, A, \{P^a_{ss'} | a \in A, s, s' \in S\})$. Note that in the discrete case there is no need to specify a $\sigma$-field, and (sub)probabilities can be specified pointwise.

### 2.5.1 Logical Characterization

The logical characterization of bisimulation metrics derives from the logical characterizations of bisimulation. We already know that, for the systems in consideration, two states are bisimilar iff they satisfy exactly the same formulas in some fixed logic (see [4], [5]). 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, so that such a formula is 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 [14]:

| 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 et al. in [4] developed a class of functional expressions, or formulas, that can be evaluated on the state space to yield values in $[0,1]$. The result of formula evaluation gives a quantitative measure of the extent to which states satisfy a particular formula. By calculating the difference of these quantities for a fixed pair of states across all formulas, a family of bisimulation metrics is constructed. Formally, we have the following:

Fix $c \in (0,1]$ and let $\mathcal{F}^c$ be a family of functional expressions whose syntax is given by the following grammar:

$$f := \mathit{1} | \max(f,f) | \langle a \rangle f | 1 - f | f \ominus q$$

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

$$\begin{aligned}
\mathit{1}(s) &= 1 \\
\max(f_1,f_2)(s) &= \max(f_1(s),f_2(s)) \\
(\langle a \rangle f)(s) &= cE_{P^a_s}[f] = c\sum_{s' \in S} P^a_{ss'} f(s') \\
(1-f)(s) &= 1 - f(s) \\
(f \ominus q)(s) &= \max(f(s) - q, 0)
\end{aligned}$$

Define $d^c : S \times S \to [0,1]$ by $d^c(s,s') = \sup_{f \in \mathcal{F}^c} |f(s) - f(s')|$. We have the following result from [4]:

**Theorem 2.5.1.** *For every $c$ in $(0,1]$, $d^c$ is a 1-bounded bisimulation metric.*

Unfortunately, while this does establish the existence of metrics possessing the desired bisimulation properties, it does not provide us with a means of computing such metrics. It is not even clear that they are computable. In fact, they are (for $c < 1$). This can be shown through the use of domain theoretical tools to establish a fixed point characterization of $\{d^c | c \in (0,1]\}$.

### 2.5.2 Fixed Point Characterization

In addition to its relational and logical characterizations, bisimulation also has a fixed point characterization. This naturally leads to the development of a class of fixed point bisimulation metrics which, as it turns out, is the same class of bisimulation metrics that result from the logic above. The main technical tool used to create this class is known as *the Knaster-Tarski Theorem for maximum fixed points* (see [28]):

**Theorem 2.5.2.** *Let $(L, \sqsubseteq)$ be a complete lattice. Let $f : L \to L$ be a monotonic function, i.e. such that if $x \sqsubseteq y$ then $f(x) \sqsubseteq f(y)$. Then $f$ has a greatest fixed point, which is also its greatest postfixed point.*[5]

Take the complete lattice $\mathcal{M}$ of 1-bounded pseudometrics and define, for a fixed $c \in (0,1]$, $F^c : \mathcal{M} \to \mathcal{M}$ by $F^c(m)(s,s') = c \max_{a \in A} m(P_s^a, P_{s'}^a)$ (here we are using the Kantorovich metric extended to subprobability distributions). Since moving metrics from states to distributions preserves the ordering of $\mathcal{M}$, it easily follows that each $F^c$ is monotonic on $\mathcal{M}$ and so has a greatest fixed point $m^c$. Adapting the proof of [5] we obtain

[5] A point $x \in L$ is said to be a fixed point of $f$ if $f(x) = x$. It is said to be postfixed if $x \sqsubseteq f(x)$.

**Theorem 2.5.3.** *For every $c$ in $(0,1]$, $m^c$ is a 1-bounded bisimulation metric.*

In fact, the following is true

**Theorem 2.5.4.** *For every $c$ in $(0,1]$, $m^c = d^c$.*

A key step in this proof relies on approximating $m^c$ via a sequence of iterates. Each $m^c$ is expressible as $\sqcap\{m_i^c\}$ where $m_0^c = \top$ and $m_{i+1}^c = F^c(m_i^c)$. Formally, one says that each $F^c$ has closure ordinal $\omega$.

An additional advantage of the existence of these iterates is that they allow one to approximate each $m^c$ to within any prescribed degree of accuracy. So each $m^c$ is computable (for $c < 1$). This follows simply by noting that by induction, $m^c(s,s') - m_i^c(s,s') \leq c^i$. That each iterate $m_i^c$ is computable follows from the computability of the Kantorovich metric.

# Chapter 3

# A Metric for Finite Markov Decision Processes

In this chapter we will emulate the creation of metrics for LMPs in order to develop such metrics for use in finite MDP compression. Since our primary concern lies with computing such metrics we will do so first via a fixed point formulation. The contents of this chapter will be, for the most part, self-contained.

## 3.1 A Metric for Markov Decision Processes

In the following we will establish a pseudometric, defined on the states of a finite MDP, satisfying the requirement that two states are assigned distance zero iff they are "behaviourally indistinguishable". We will formally specify what we mean by "indistinguishable" in a moment. First, we introduce the model and assumptions to be used throughout.

Let $M = (S, A, \{P^a_{ss'} : a \in A, s, s' \in S\}, \{r^a_s : a \in A, s \in S\})$ be a finite MDP where:

- $A$ is a finite set of actions such that $\forall s \in S, A_s = A$. We will denote the cardinality of $A$ by $|A|$ and sometimes enumerate $A$ as $\{a_1, \ldots, a_{|A|}\}$.
- Since rewards are bounded we will assume without loss of generality that $\forall a \in A, \forall s \in S, r^a_s \in [0, 1]$.

### 3.1.1 Bisimulation

When are two states indistinguishable? In [8], Givan et al. examined different notions of state equivalence to answer this very question. They concluded that the most appropriate notion, derived from the theory of concurrent processes, is *stochastic bisimulation.*

**Definition 3.1.1.** A *(stochastic) bisimulation relation* $R$ is an equivalence relation on $S$ that satisfies the following property:

$$sRs' \Leftrightarrow \forall a \in A, (r^a_s = r^a_{s'} \text{ and } \forall C \in S/R, P^a_s(C) = P^a_{s'}(C))$$

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

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

Roughly speaking, two states $s$ and $s'$ bisimilar iff 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 for MDPs can also be formulated via fixed point theory and it will be fruitful for our purposes to do so.

**Definition 3.1.2.** Let $(Rel, \subseteq)$ be the complete lattice of binary relations on $S$. Define $F : Rel \to Rel$ by

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

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

Note that the existence of a greatest fixed point in the definition above is guaranteed by the Knaster-Tarski Fixed Point Theorem since $F$ is monotone on $Rel$.

**Lemma 3.1.3.** *The definitions of bisimulation in 3.1.1 and 3.1.2 are equivalent.*

*Proof.* Clearly, the greatest fixed point of $F$ is a bisimulation relation and therefore contained in $\sim$, the largest bisimulation relation. On the other hand, $\sim$ is easily seen to be a fixed point of $F$ and is therefore contained in its greatest fixed point. □

Unfortunately, bisimulation is too restrictive. Consider the MDP in figure 3.1 with 4 states labeled $s$, $t$, $u$, and $v$, and 1 action labeled $a$. Suppose:

- $r_v^a = 0$. Then all states share the same immediate reward. Moreover, starting in any of the four states one transitions to one of the four states with probability one. Thus, all states can be grouped together in one bisimulation class; that is, all states are bisimilar.

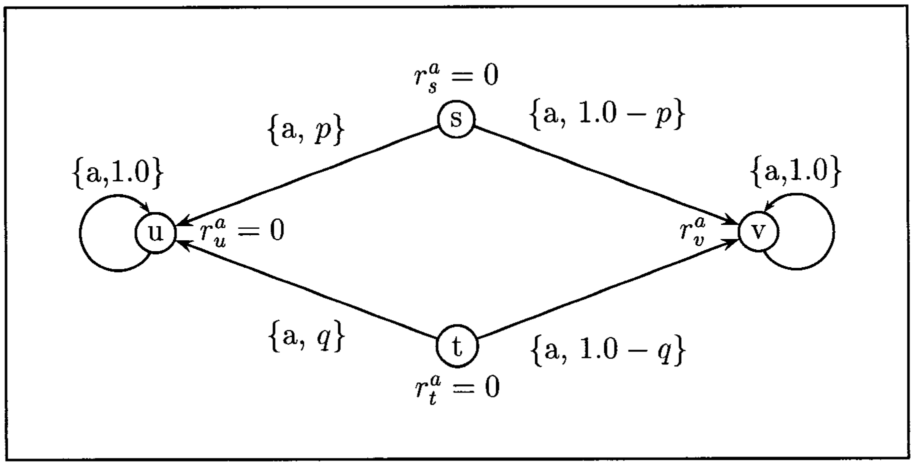

Figure 3.1: If $r_v^a = 0$ then all states are bisimilar. On the other hand, if $r_v^a > 0$ then $s \sim t \iff p = q$, $s \sim u \iff p = 1$ and $t \sim u \iff q = 1$.

- $r_v^a > 0$. Then $v$ is the only state in its bisimulation class since it is the only one with a positive reward. Moreover, $s$ and $t$ are bisimilar iff they share the same probability of transitioning to $v$'s bisimulation class. Each is bisimilar to $u$ iff that probability is zero. Thus, $u$, $s$, $t$ $\not\sim v$, $s \sim t \iff p = q$, $s \sim u \iff p = 1.0$, and $t \sim u \iff q = 1.0$.

This demonstrates that bisimulation alone is simply too strong a notion. If $r_v$ is just slightly positive, and $p$ differs only slightly from $q$ we should expect $s$ and $t$ to be practically bisimilar. However, such a fine distinction cannot be made; all we can say is that two states are bisimilar or they are not.

This is where metrics come to the rescue. They allow us to give a quantitative notion of bisimulation, susceptible to slight variations in the parameters of the model. We will soon show how we can associate "how bisimilar" a pair of states are to a distance between zero and one.

### 3.1.2 A Class of Bisimulation Metrics

Recall that $\mathcal{M}$ is the space of 1-bounded pseudometrics on $S$, partially ordered as follows:

$$m_1 \preceq m_2 \Leftrightarrow \forall s, s' \in S, m_1(s, s') \geq m_2(s, s')$$

We say $m \in \mathcal{M}$ is a *bisimulation metric* iff $m(s, s') = 0 \Leftrightarrow s \sim s'$. As previously mentioned, our goal is to establish such metrics for finite MDP state aggregation in continual RL tasks.

One of the first things we can remark about bisimulation metrics is that they constitute a class of equivalent metrics. Let $bis$ be the bisimulation metric which assigns distance 1 to all pairs of non-bisimilar states. Then for any bisimulation metric $m$, $m(s, s') \leq bis(s, s')$ for all $s, s'$. On the other hand, let $c_{min} = \min m(s, s')$ where the minimum is taken over all $s, s'$ where $m$ is positive (of course if no such states exist then $\top$, the everywhere zero metric, is the only bisimulation metric for this particular model). Then $c_{min} > 0$ and for every $s, s' \in S$, $c_{min} bis(s, s') \leq m(s, s')$. Thus, all bisimulation metrics are equivalent to $bis$, and to each other. Moreover, it is immediate that any metric equivalent to a bisimulation metric must itself be a bisimulation metric.

What else can we say about bisimulation metrics? Consider a collection of metrics, $\{d_a\}_{a \in A} \subseteq \mathcal{M}$, satisfying

$$d_a(s, s') = 0 \Leftrightarrow r_s^a = r_{s'}^a$$

(e.g., take $d_a(s, s') = |r_s^a - r_{s'}^a|$). Then we have:

**Lemma 3.1.4.** *If $m$ is a bisimulation metric then*

$$\forall s, s' \in S, m(s,s') = 0 \Leftrightarrow \forall a \in A, (d_a(s,s') = 0 \text{ and } m(P_s^a, P_{s'}^a) = 0) \quad (3.1)$$

*Proof.* First note that for a bisimulation metric we may rewrite the primal LP for $m(P_s^a, P_{s'}^a)$ as

$$\max_{u_C} \sum_{C \in S/\sim} (P_s^a(C) - P_{s'}^a(C)) u_C$$

$$\text{subject to: } \forall C, D, u_C - u_D \leq \min_{i \in C, j \in D} m(s_i, s_j)$$

$$\forall C, 0 \leq u_C \leq 1$$

and so the forward direction is immediate.

For the converse, note that if $m(P_s^a, P_{s'}^a) = 0$ then $P_s^a(C) = P_{s'}^a(C)$ for every equivalence class $C$. Suppose to the contrary that $\exists C$ such that $P_s^a(C) \neq P_{s'}^a(C)$. WLOG $P_s^a(C) > P_{s'}^a(C)$. Clearly $C \neq S$, so we may take $u_C = \min_{k \in C, j \in D} m(s_k, s_j)$ and $u_D = 0$ for all other classes and obtain a contradiction by way of a positive lower bound on $m(P_s^a, P_t^a)$. The result now follows. □

Is condition 3.1 sufficient as well? It seems reasonable to expect that any metric that assigns distance zero to a pair of states if and only if it assigns distance zero to both their rewards and their distributions should be a bisimulation metric. This is easily seen to not be the case, as the metric $\perp$ satisfies the condition and is not necessarily a bisimulation metric, as illustrated by the MDP in figure 3.2. However, we do have the following:

**Lemma 3.1.5.** *Suppose $m \in \mathcal{M}$ satisfies condition 3.1. Then*

$$m(s, s') = 0 \Rightarrow s \sim s'$$

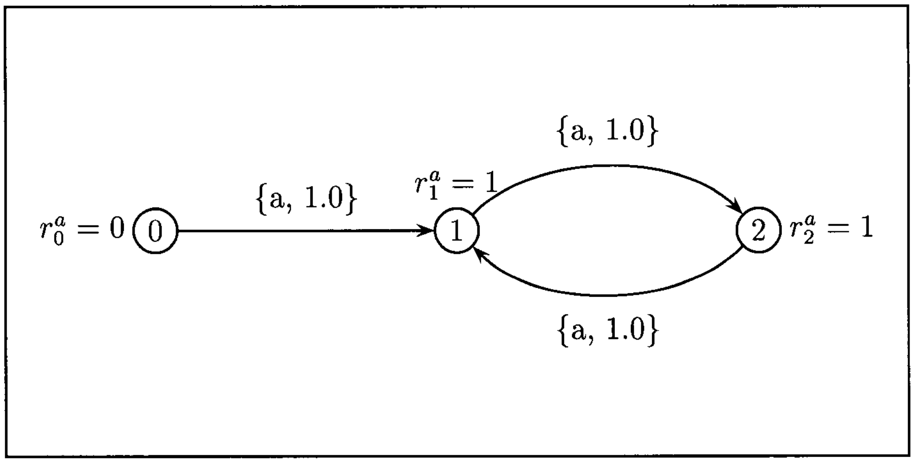

Figure 3.2: $\perp(P^a_s, P^a_{s'}) = 1$ for $s \neq s'$, so that $\perp$ satisfies the bisimulation metric condition. However, it is not a bisimulation metric since it assigns distance 1 to bisimilar states 1 and 2.

*Proof.* Define a relation $R$ on $S$ by $sRs' \Leftrightarrow m(s, s') = 0$. We will show that $R$ is a bisimulation relation. It is clearly an equivalence relation. Now suppose $sRs'$. Then $\forall a \in A, r^a_s = r^a_{s'}$ and $m(P^a_s, P^a_{s'}) = 0$. Just as in the previous lemma, we may rewrite the primal LP as

$$\max_{u_C} \sum_{C \in S/R} (P^a_s(C) - P^a_{s'}(C))u_C$$

$$\text{subject to: } \forall C, D, u_C - u_D \leq \min_{i \in C, j \in D} m(s_i, s_j)$$

$$\forall C, 0 \leq u_C \leq 1$$

so that $P^a_s(C) = P^a_{s'}(C)$, as well. Thus, $s \sim s'$. □

Although condition 3.1 is not sufficient it does indicate the strong relationship between the distance a bisimulation metric assigns to a pair of states and the distance it assigns to the states' induced distributions, as well as the distance between the states' immediate rewards. Motivated by this, we will construct bisimulation metrics of the following form

$$m(s,s') = \underset{a\in A}{\phi}\left(d_a(s,s'), m(P_s^a, P_{s'}^a)\right)$$

The underlying functions we will utilize are as follows:
Let $\phi : [0,1]^{2|A|} \rightarrow [0,1]$

1. $\phi(\vec{x}) = 0$ iff $\vec{x} = \vec{0}$.

2. $\forall \vec{x},\ \vec{y},\ \vec{z} \in [0,1]^{2A}, \vec{z} \leq \vec{x} + \vec{y} \Rightarrow \phi(\vec{z}) \leq \phi(\vec{x}) + \phi(\vec{y})$

3. $\phi$ is continuous coordinate-wise, i.e. for every $i$ between 1 and $2|A|$

$$\bigsqcup_n \phi(x_1, \ldots, x_i^n, \ldots, x_{2|A|}) = \phi(x_1, \ldots, \bigsqcup_n x_i^n, \ldots, x_{2|A|})$$

We are now ready to construct our bisimulation metrics. We do so in a manner analogous to how we constructed the maximum bisimulation relation. It is a somewhat surprising that in order to do so, our underlying functions need only satisfy the first two properties above.

**Theorem 3.1.6.** *Let $(\mathcal{M}, \preceq)$ be the complete lattice of 1-bounded pseudometrics on $S$. Let $\phi : [0,1]^{2|A|} \rightarrow [0,1]$ satisfy 1 and 2 above.*
*Define $F^\phi : \mathcal{M} \rightarrow \mathcal{M}$ by*

$$F^\phi(m)(s,s') = \phi(x_1, \ldots, x_{2|A|}) \textit{ where}$$
$$x_i = d_{a_i}(s,s') \textit{ and}$$
$$x_{A+i} = m(P_s^{a_i}, P_{s'}^{a_i}) \textit{ for } 1 \leq i \leq |A|.$$

*Then $s$ and $s'$ are bisimilar iff $m^{\phi}(s, s') = 0$ where $m^{\phi}$ is the greatest fixed point of $F^{\phi}$.*

*Proof.* We have three things to show; namely, $\forall m \in \mathcal{M}.\ F^{\phi} \in \mathcal{M}$ (i.e. $F^{\phi}$ is well-defined), $F^{\phi}$ has a greatest fixed point $m^{\phi}$, and $m^{\phi}$ is a bisimulation metric.

Let $m$ be in $\mathcal{M}$. We need to prove that $F^{\phi}(m)$ is a 1-bounded pseudometric. It is clearly non-negative and 1-bounded. $F^{\phi}(m)(s, s) = 0$ easily follows from $m$ and $d_a$ being pseudometrics and $\phi$ taking value zero at the origin. Symmetry also easily follows. For the triangle inequality, we first recall that $m$ applied to distributions also satisfies a triangle inequality. So, by property 2, we have

$$\begin{aligned} F^{\phi}(m)(s, s') &\leq \phi(d_{a_1}(s, u), \ldots, m(P_s^{a_{2|A|}}, P_u^{a_{2|A|}})) \\ &\qquad + \phi(d_{a_1}(u, s'), \ldots, m(P_u^{a_{2|A|}}, P_{s'}^{a_{2|A|}})) \\ &= F^{\phi}(m)(s, u) + F^{\phi}(m)(u, s') \end{aligned}$$

So $F^{\phi}$ is well-defined.

To establish the existence of a greatest fixed point of $F^{\phi}$ we appeal to the Knaster-Tarski Fixed Point Theorem. So we need only show that $F^{\phi}$ is monotone on $\mathcal{M}$. Note that properties 1 and 2 above imply that each underlying $\phi$ is monotone. Since in applying metrics to distributions the ordering of metrics is preserved, monotonicity of $F^{\phi}$ clearly holds. Thus, $m^{\phi}$ exists for each $\phi$.

Finally, to establish that $m^{\phi}$ is a bisimulation metric first note that by construction it satisfies condition 3.1. So by the preceding lemma, $m^{\phi}(s, s') = 0 \Rightarrow s \sim s'$.

For the other direction, recall that $bis$ is the bisimulation metric which assigns 1 to all pairs of non-bisimilar states. By property 1 and lemma 3.1.4, $F^{\phi}(bis)(s,t) = 0 \Leftrightarrow \forall a \in A, (d_a(s,s') = 0 \text{ and } bis(P_s^a, P_{s'}^a) = 0) \Leftrightarrow bis(s,s') = 0$. It follows that $bis \preceq F^{\phi}(bis)$, i.e. $bis$ is a postfixed point of $F^{\phi}$. Since $m^{\phi}$ is also the greatest postfixed point of $F^{\phi}$, $b \preceq m^{\phi}$. Thus, $s \sim s' \Rightarrow m^{\phi}(s,s') = 0$. □

### 3.1.3 A Class of RL Metrics

In the last section, we established a class of metrics each of which agrees with bisimulation. At this point it is necessary to recall exactly why we did so. We seek to aggregate large state space MDPs in order to solve continual RL tasks. The correct notion for MDP state aggregation is bisimulation, but it is too strong. Thus, we appeal to metrics to give us a quantitative form of bisimulation. Specifically, we want a bisimulation metric as previously defined. However, we also desire that such a metric should reflect the variations in the difference between rewards and the difference between distributions. Moreover, we would like to recover information useful for continual RL tasks. In [8], Givan et al. were able to show that bisimilar states have the same optimal value. A desired property for our metrics is that if two states are close together under a bisimulation metric, then their optimal values should also be close together.

Now we have managed to establish a class of bisimulation metrics, but what of the other desired characteristics? Unfortunately, they can fail to hold for even the simplest bisimulation metrics. Consider the underlying function that attains value 1 everywhere but at the origin. It is not hard to see that the resulting bisimulation metric is $bis$. The problem here is that the ability of $bis$ to distinguish states is exactly the same as that of bisimulation; it is too strong. There is no quantitative difference between pairs of states that are almost bisimilar and those that are really different. What we need to do is restrict our attention to a subclass of bisimulation metrics possessing the desired properties.

For inspiration we look to the Bellman Equations for the optimal value function, which yield the following bound:

$$|V^*(s) - V^*(s')| \leq \max_{a \in A}(|r_s^a - r_{s'}^a| + \gamma | \sum_{u \in S} (P_{su}^a - P_{s'u}^a) V^*(u)|)$$

The first component of the RHS is simply the distance in immediate rewards, while the second component is strikingly similar to the primal LP for the distance in distributions.

Based on these observations we fix a particular class of underlying functions, $\{\phi(\vec{x}) = \max_{1 \leq i \leq A}(c_R x_i + c_T x_{A+i})\}$, indexed by two positive constants, $c_R$ and $c_T$, in our previous theorem to obtain the class of RL metrics:

**Corollary 3.1.7.** *Let $c_R$, $c_T \in (0,1)$ s.t. $c_R + c_t \leq 1$. Define $F^{c_R,c_T} : \mathcal{M} \to \mathcal{M}$ by*

$$F^{c_R,c_T}(m)(s,s') = \max_{a \in A}(c_R d_a(s,s') + c_T m(P_s^a, P_{s'}^a))$$
$$\textit{where } d_a(s,s') = |r_s^a - r_{s'}^a|.$$

*Then the RL metric $m^{c_R,c_T}$ is the greatest fixed point of $F^{c_R,c_T}$.*

The constants $c_R$ and $c_T$ weight the distance between rewards and the distance between transition distributions respectively.

It is clear from the construction that such metrics will reflect the variations in rewards and distributions; but, do they provide the aforementioned bounds on optimal values? Indeed, they do; but before we can show that we need an alternate formulation of the RL metrics.

### 3.1.4 An Alternate Characterization of RL Metrics

Note that the underlying function for each RL metric

$$\phi(\vec{x}) = \max_{1 \leq i \leq A} (c_R x_i + c_T x_{A+i})$$

satisfies properties 1,2,3 above. For such functions $\phi$, $m^\phi$ can be expressed as the limit of a sequence of iterates.

**Lemma 3.1.8.** *Let $\phi$ satisfy properties 1-3. Then $F^\phi$ has closure ordinal $\omega$, i.e. $m^\phi = \sqcap m_i^\phi$ where $m_0^\phi = \top$ and $m_{i+1}^\phi = F^\phi(m_i^\phi)$.*

*Proof.* Since $m^\phi \preceq \top$ and $F^\phi$ is monotone we see by induction that $m^\phi \preceq \sqcap\, m_i^\phi$. All that remains is to show that $\sqcap m_i^\phi$ is a fixed point of $F^\phi$. Simply note that for all $(s, s')$

$$\begin{aligned}
&(\sqcap\{m_i^\phi\})(s,s') \\
&= \sqcup m_i^\phi(s,s') \\
&= \sqcup\{\sqcup m_{i+1}^\phi(s,s'), 0\} \\
&= \sqcup m_{i+1}^\phi(s,s') \\
&= \sqcup F^\phi(m_i^\phi)(s,s') \\
&= \sqcup \phi(d_{a_1}(s,s'), \ldots, d_{a_{|A|}}(s,s'), m_i^\phi(P_s^{a_1}, P_{s'}^{a_1}), \ldots, m_i^\phi(P_s^{a_{|A|}}, P_{s'}^{a_{|A|}}))
\end{aligned}$$

Since all the coordinate values, that is the $d_{a_k}(s,s')$'s and the $m_i^\phi(P_s^{a_k}, P_{s'}^{a_k})$'s, are nondecreasing with respect to $i$, and $\phi$ is monotone, we may replace the single supremum over $i$ by a supremum over all coordinates (specifically, over the last $|A|$ coordinates, since the first $|A|$ coordinate values are constant with respect to $i$). This allows us to use coordinate-wise continuity (property 3 of $\phi$) to move the suprema "inside" $\phi$ and apply lemma 2.4.7 to the appropriate coordinate values:

$$
\begin{aligned}
&= \sqcup_{i_1,\ldots,i_{|A|}} \phi(d_{a_1}(s,s'),\ldots,d_{a_{|A|}}(s,s'), m_{i_1}^\phi(P_s^{a_1}, P_{s'}^{a_1}),\ldots, m_{i_{|A|}}^\phi(P_s^{a_{|A|}}, P_{s'}^{a_{|A|}}))\\
&= \phi(d_{a_1}(s,s'),\ldots,d_{a_{|A|}}(s,s'), \sqcup_{i_1} m_{i_1}^\phi(P_s^{a_1}, P_{s'}^{a_1}),\ldots, \sqcup_{i_{|A|}} m_{i_{|A|}}^\phi(P_s^{a_{|A|}}, P_{s'}^{a_{|A|}}))\\
&= \phi(d_{a_1}(s,s'),\ldots,d_{a_{|A|}}(s,s'), (\sqcap_{i_1} m_{i_1}^\phi)(P_s^{a_1}, P_{s'}^{a_1}),\ldots, (\sqcap_{i_{|A|}} m_{i_{|A|}}^\phi)(P_s^{a_{|A|}}, P_{s'}^{a_{|A|}}))\\
&= F^\phi(\sqcap\{m_i^\phi\})(s,s')
\end{aligned}
$$

□

In addition, each RL metric can be expressed via a real valued modal logic. The intuition is exactly the same as that for LMP bisimulation metrics: the closer two states are (with respect to bisimulation), the greater is the complexity of the simplest distinguishing formula.

### 3.1.5 A Logical Characterization

Let $\mathcal{F}^{c_R,c_T}$ be the family of functions whose syntax is given by:

$$f := 1 | \max(f,f) | h \circ f | \langle a \rangle f$$

where $a \in A$ and $h$ is non-expansive on $[0,1]$. These function expressions are evaluated on $S$ as follows:

$$\begin{aligned} \mathit{1}(s) &= 1 \\ \max(f_1, f_2)(s) &= \max(f_1(s), f_2(s)) \\ (h \circ f)(s) &= h(f(s)) \\ (\langle a \rangle f)(s) &= c_R r_s^a + c_T E_{P_s^a}[f] \end{aligned}$$

**Theorem 3.1.9.** *Let $m^{c_R,c_T}$ be the RL metric obtained via corollary 3.1.7. Then $m^{c_R,c_T}(s,s') = \sup_{f \in \mathcal{F}^{c_R,c_T}} |f(s) - f(s')|$*

The proof of this theorem is adapted from [5]. For convenience we will work with a fixed pair of weights, $c_R$ and $c_T$, and omit these as superscripts from $\mathcal{F}^{c_R,c_T}$, $F^{c_R,c_T}$, $m^{c_R,c_T}$, and $\{m_i^{c_R,c_T}\}$ in all that follows.

Define $d : S \times S \to [0,1]$ by $d(s,s') = \sup_{f \in \mathcal{F}} |f(s) - f(s')|$. In order to prove theorem 3.1.9 we first need to establish some intermediate technical results. The first formalizes what is intuitively clear from observing the four logical expressions defining the logic: the modal operator, $\langle a \rangle f$, alone makes distances bigger; applying a non-expansive function or a taking the maximum only decreases the distances between states.

**Lemma 3.1.10.** *For every pair of states $s, s'$,*

$$d(s,s') = \sup_{a \in A, f \in \mathcal{F}} |(\langle a \rangle f)(s) - (\langle a \rangle f)(s')|$$

*Proof.* Clearly, $d(s,s') \geq \sup_{a \in A, f \in \mathcal{F}} |(\langle a \rangle f)(s) - (\langle a \rangle f)(s')|$. For the other direction we proceed by structural induction. Specifically, we show

$$\forall f \in \mathcal{F},\ \exists f' \in \mathcal{F} : |f(s) - f(s')| \leq |(\langle a \rangle f')(s) - (\langle a \rangle f')(s')|$$

This is trivial in the cases $f = \mathit{1}$ or $f = \langle a \rangle f'$. If $f = \max(f_1, f_2)$ or $f = h \circ f'$ then note that

$$|\max(f_1, f_2)(s) - \max(f_1, f_2)(s')| \leq \max_{i \in \{1,2\}} |f_i(s) - f_i(s')|$$

and

$$|(h \circ f')(s) - (h \circ f')(s')| \leq |f'(s) - f'(s')|$$

so that the result follows by the induction hypothesis. □

The next result shows us that the Kantorovich metric with the logical distance, $d$, admits an explicit formulation solely in terms of the logic.

**Lemma 3.1.11.** *Let $P$ and $Q$ be probability distributions on $S$. Then*

$$d(P, Q) = \sup_{f \in \mathcal{F}} |E_P[f] - E_Q[f]|$$

*Proof.* Let $f \in \mathcal{F}$ and define $u_i = f(s_i)$. Then $0 \leq u_i \leq 1$ and $u_i - u_j \leq d(s_i, s_j)$ so that $\{u_i\}$ constitutes a feasible solution to the primal LP for $d(P, Q)$. Since $\forall f \in \mathcal{F}, |E_P[f] - E_Q[f]| \leq d(P, Q)$, it follows that $d(P, Q) \geq \sup_{f \in \mathcal{F}} |E_P[f] - E_Q[f]|$.

For the other direction we first remark that the following are non-expansive functions on $[0, 1]$: For $q \in [0, 1]$ let

$$h_q(x) = \begin{cases} 0 & \text{if } 0 \leq x \leq q \\ x - q & \text{otherwise} \end{cases}, k_q(x) = \begin{cases} x + q & \text{if } 0 \leq x \leq 1 - q \\ 1 & \text{otherwise} \end{cases}$$

(we may rewrite these in terms of the logic given for LMPs as $h_q(x) = x \ominus q$ and $k_q(x) = 1 - ((1 - x) \ominus q))$

Now, let $\{u_i\}$ maximize $d(P, Q)$ and let $\epsilon > 0$. We will first show

$$\forall i, j,\ \exists f_{ij} \in \mathcal{F} : f_{ij}(s_i) = u_i \text{ and } f_{ij}(s_j) \leq \max(0, u_i - d(s_i, s_j) + \epsilon).$$

To see this note that

$$d(s_i, s_j) - \epsilon < d(s_i, s_j) \Rightarrow \exists h_{ij} \in \mathcal{F} : d(s_i, s_j) - \epsilon < |h_{ij}(s_i) - h_{ij}(s_j)|$$

WLOG $|h_{ij}(s_i) - h_{ij}(s_j)| = h_{ij}(s_i) - h_{ij}(s_j)$ so that $h_{ij}(s_j) - h_{ij}(s_i) < -d(s_i, s_j) + \epsilon$. Suppose:

- $h_{ij}(s_i) \leq u_i$. Take $q = u_i - h_{ij}(s_i) \in [0, 1]$ and define $f_{ij} = k_q \circ h_{ij} \in \mathcal{F}$.
- $h_{ij}(s_i) > u_i$. Take $q = h_{ij}(s_i) - u_i \in [0, 1]$ and define $f_{ij} = h_q \circ h_{ij} \in \mathcal{F}$.

In both cases it is clear that $f_{ij}$ satisfies the required properties.

Now take $f_i = \min_j f_{ij} \in \mathcal{F}$. Then $f_i(s_i) = u_i$ and for $j \neq i$, $f_i(s_j) \leq f_{ij}(s_j) \leq u_j + \epsilon$.

Finally, take $f = \max_i f_i \in \mathcal{F}$. Then $u_j \leq f(s_j) \leq u_j + \epsilon$. Thus,

$$\begin{aligned} d(P, Q) &\leq (E_P[f] - E_Q[f]) + \epsilon \sum_{i:P(s_i)<Q(s_i)} (Q(s_i) - P(s_i)) \\ &\leq \sup_{f \in \mathcal{F}} |E_P[f] - E_Q[f]| + \epsilon \sum_{i:P(s_i)<Q(s_i)} (Q(s_i) - P(s_i)) \end{aligned}$$

Since $\sum_{i:P(s_i)<Q(s_i)}(Q(s_i) - P(s_i)) \geq 0$ and $\epsilon$ is arbitrary, it follows that $d(P, Q) \leq \sup_{f \in \mathcal{F}} |E_P[f] - E_Q[f]|$. □

Recall that $m$ is the greatest fixed point of $F$. Thus, to prove theorem 3.1.9 we need only verify two facts: $d$ is a fixed point of $F$, and $d$ is at least as big as $m$ (with respect to the ordering of metrics). The first fact is established through the preceding lemmas.

**Lemma 3.1.12.** *Let $F$ be the functional defined in corollary 3.1.7. Then the logical metric,d, is a fixed point of $F$.*

*Proof.* By lemma 3.1.10 and lemma 3.1.11 it follows that:

$$\begin{aligned}
d(s,s') &= \sup_{a\in A, f\in\mathcal{F}} |((\langle a\rangle f)(s) - ((\langle a\rangle f)(s')| \\
&\leq \sup_{a\in A, f\in\mathcal{F}} (c_R|r_s^a - r_{s'}^a| + c_T|E_{P_s^a}[f] - E_{P_{s'}^a}[f]|) \\
&= \max_{a\in A}(c_R d_a(s,s') + c_T \sup_{f\in\mathcal{F}} |E_{P_s^a}[f] - E_{P_{s'}^a}[f]|) \\
&= \max_{a\in A}(c_R d_a(s,t) + c_T d(P_s^a, P_{s'}^a) \\
&= F(d)(s,s')
\end{aligned}$$

So $F(d) \preceq d$.

Let $\epsilon > 0$. Then $\exists f \in \mathcal{F},\ a \in A$ such that

$$F(d)(s,s') - \epsilon < c_R d_a(s,s') + c_T|E_{P_s^a}[f] - E_{P_{s'}^a}[f]|$$

WLOG $d_a(s,s') = r_s^a - r_{s'}^a$. Take $f' \in \mathcal{F}$ to be:

- $\langle a\rangle f$ if $E_{P_s^a}[f] \geq E_{P_{s'}^a}[f]$
- $\langle a\rangle(1-f)$ if $E_{P_s^a}[f] < E_{P_{s'}^a}[f]$

Then

$$F(d)(s,s') - \epsilon < |f'(s) - f'(s')| \leq d(s,s')$$

Since $\epsilon$ is arbitrary, it follows that $d \preceq F(d)$. □

We are now ready to prove theorem 3.1.9.

*Proof.* Given the preceding lemmas, we now need only show that $d$ is the greatest fixed point of $F$. Clearly $d \preceq m$.

For the other inequality, we first define the depth, $l(\cdot)$, of a function in $\mathcal{F}$. Define $l : \mathcal{F} \to \mathbb{N}$ inductively by

$$\begin{aligned} l(\mathit{1}) &= 0 \\ l(\max(f_1, f_2)) &= \max(l(f_1), l(f_2)) \\ l(h \circ f) &= l(f) \\ l(\langle a \rangle f) &= 1 + l(f) \end{aligned}$$

Let $\mathcal{F}_i = \{f \in \mathcal{F} | l(f) \leq i\}$. Then it immediately follows that:

- $\mathcal{F}_i \uparrow \mathcal{F}$
- if we define $d_i$ by $d_i(s, s') = \sup_{f \in \mathcal{F}_i} |f(s) - f(s')|$ then $d_i \downarrow d$

We will show by induction that $\forall i \geq 0, m_i \preceq d_i$.

For $i = 0$, $f \in \mathcal{F}_i$ must be one of $\mathit{1}$, $max(f_1, f_2)$, or $h \circ f'$. An easy structural induction shows that $|f(s) - f(s')| \leq 0$, so that $m_0 \preceq d_0$.

For $i + 1$ we again use structural induction, this time to show that for $f \in \mathcal{F}_{i+1}, |f(s) - f(s')| \leq m_{i+1}(s, s')$. The only interesting case is $f = \langle a \rangle f'$ where $f' \in \mathcal{F}_i$. In this case, note that by the induction hypothesis (over $i$)

$$f'(s_k) - f'(s_j) \leq d_i(s_k, s_j) \leq m_i(s_k, s_j)$$

so that $\{f'(s_k)\}$ constitutes a feasible solution to the primal LP for $m_i(P_s^a, P_{s'}^a)$.

Therefore,

$$\begin{aligned}
|f(s) - f(s')| &\leq \max_{a\in A} |((\langle a\rangle f')(s) - ((\langle a\rangle f')(s')| \\
&\leq \max_{a\in A}(c_R d_a(s,s') + c_T |E_{P^a_s}[f'] - E_{P^a_{s'}}[f']|) \\
&\leq \max_{a\in A}(c_R d_a(s,s') + c_T m_i(P^a_s, P^a_{s'})) \\
&\leq m_{i+1}(s,s')
\end{aligned}$$

so that $d_{i+1}(s,s') \leq m_{i+1}(s,s')$.

Finally, taking limits yields $m \preceq d$. □

Let use reiterate the significance of the preceding result. The RL metric, $m$, obtained as a fixed point in corollary 3.1.7 can be equivalently expressed in terms of a real valued modal logic. This logical formulation provides the following intuitive reasoning behind the distances assigned by $m$: the distance assigned to states is related to the quantitative difference in the formulas they satisfy. Thus, $m$ is a quantitative generalization of the exact matching of logical properties of states. It follows that $m$ could be used to potentially analyze quantitatively certain logical properties of MDPs. Moreover, the family of logical expressions potentially provides a means of computing the distances assigned by $m$. [1]

## 3.2 Bounds for the Optimal Value Function

Recall that if we define $\forall s \in S.\ V_0(s) = 0$ and

$$V_{i+1}(s) = \max_{a\in A}(r^a_s + \gamma \sum_{u\in S} P^a_{su} V_i(u))$$

[1] A more detailed discussion of metric computability will be carried out at the end of this chapter.

then $V_i$ converges to $V^*$, the optimal value function with discount factor $\gamma \in [0, 1)$.

### 3.2.1 The Original MDP

We have previously stated that it is reasonable to expect that if two states are close together according to the RL metric then their optimal values should be close too. The following bounds formally establish that this is so.

**Theorem 3.2.1.** *Suppose $\gamma \leq c_T$. Then $\forall s, s' \in S$:*

1. $|V_i(s) - V_i(s')| \leq \frac{m_i(s,s')}{c_R}$, $\forall i \geq 0$.
2. $|V^*(s) - V^*(s')| \leq \frac{m(s,s')}{c_R}$

*Proof.* Clearly the proof of the second item follows from the first by taking limits. For the proof of the first item we proceed by induction. The inequality holds trivially for $i = 0$. Now note that since $\gamma \leq c_T$

$$0 \leq \frac{c_R \gamma}{c_T} V_i(u) \leq \frac{(1 - c_T)\gamma}{c_T(1 - \gamma)} \leq 1$$

and by the induction hypothesis

$$\frac{c_R \gamma}{c_T} V_i(u) - \frac{c_R \gamma}{c_T} V_i(v) \leq c_R |V_i(u) - V_i(v)| \leq m_i(u, v).$$

So $\{\frac{c_R \gamma}{c_T} V_i(u) : u \in S\}$ constitutes a feasible solution for the primal LP for

$m_i(P_s^a, P_{s'}^a)$. It follows that

$$\begin{aligned}
c_R|V_{i+1}(s) - V_{i+1}(s')| &= c_R|\max_{a\in A}(r_s^a + \gamma\sum_{u\in S} P_{su}^a V_i(u)) - \max_{a\in A}(r_{s'}^a + \gamma\sum_{u\in S} P_{s'u}^a V_i(u))| \\
&\leq c_R \max_{a\in A}|r_s^a - r_{s'}^a + \gamma\sum_{u\in S}(P_{su}^a - P_{s'u}^a)V_i(u)| \\
&\leq \max_{a\in A}(c_R|r_s^a - r_{s'}^a| + c_T|\sum_{u\in S}(P_{su}^a - P_{s'u}^a)(\frac{c_R\gamma}{c_T}V_i(u))|) \\
&\leq \max_{a\in A}(c_R|r_s^a - r_{s'}^a| + c_T m_i(P_s^a, P_{s'}^a)) \\
&= F(m_i)(s, s') \\
&= m_{i+1}(s, s')
\end{aligned}$$

□

### 3.2.2 The Aggregate MDP

Now suppose that we aggregate the state space of our original MDP $M$ and solve the RL problem on the resultant MDP. We would hope that optimal values assigned to states belonging to the same class would be related somehow to the optimal value assigned to that class. In the following we will establish such a relationship, one that shows that if the states in a particular class are close together according to the RL metric then so are the optimal values of these states and the optimal value of the class.

First, we fix some notation and assumptions concerning the aggregated MDP $M'$:

$$M' = (S', A, \{P_{CD}^a : a \in A, C, D \in S'\}, \{r_C^a : a \in A, C \in S'\})$$

where:

- $S'$ is a partition of the state space of $M$. That is, it is a collection of disjoint nonempty subsets of $S$ whose union is $S$. Since $S$ is finite, so is $S'$. We will occasionally denote the class containing state $s \in S$ by $C_s$.

- $A$ is the same finite set of actions as in $M$.

- $P^a_{CD}$ is the probability of transitioning from class $C$ to class $D$ under action $a$. The probability distribution induced by class $C$ and action $a$ on $S'$ is denoted by $P^a_C$. We assume this to be defined as the average probability of transitioning to $D$, where the average is taken over all states in $C$. That is,

$$P^a_{CD} = \frac{1}{|C|} \sum_{s \in C, s' \in D} P^a_{ss'}$$

  Clearly each is non-negative. Moreover, its total mass is given by

$$\begin{aligned}\sum_{D \in S'} P^a_{CD} &= \frac{1}{|C|} \sum_{s \in C} \sum_{D \in S'} \sum_{s' \in D} P^a_{ss'} \\ &= \frac{1}{|C|} \sum_{s \in C} \sum_{s' \in S} P^a_{ss'} \\ &= \frac{1}{|C|} \sum_{s \in C} 1 \\ &= 1\end{aligned}$$

  So that each $P^a_C$ is indeed a probability distribution on $S'$ for each $a$.

- $r^a_C$ is the reward associated with taking action $a$ in class $C$. We assume that each is the average of the rewards received in that class for that action, i.e.

$$r_C^a = \frac{1}{|C|}\sum_{s\in C} r_s^a$$

We are now ready to describe the relationship between the optimal values of $s$, $C_s$, and the RL metric $m$. An important quantity in the following will be the average distance under metric $m$ from $s$ to all states in $C_s$,

$$\frac{1}{|C_s|}\sum_{s'\in C_s} m(s,s')$$

which for purposes of convenience will be denoted by $avg(s,m)$.

**Theorem 3.2.2.** *Suppose $\gamma \le c_T$. Then $\forall s \in S$:*

1. $|V_i(C_s) - V_i(s)| \le \frac{1}{c_R}(avg(s,m_i) + \sum_{k=1}^{i-1} \gamma^{i-k} \max_{u\in S} avg(u,m_k))$, $\forall i \ge 0$

2. $|V^*(C_s) - V^*(s)| \le \frac{1}{c_R}(avg(s,m) + \frac{\gamma}{1-\gamma}\max_{u\in S} avg(u,m))$

*Proof.* Once more we proceed by induction. The inequality holds trivially for $i = 0$.

$$
\begin{aligned}
&|V_{i+1}(C_s) - V_{i+1}(s)| \\
&= |\max_{a\in A}(r_{C_s}^a + \gamma\sum_{D\in S'} P_{C_sD}^a V_i(D)) - \max_{a\in A}(r_s^a + \gamma\sum_{u\in S} P_{su}^a V_i(u))| \\
&\le \frac{1}{|C_s|}\sum_{s'\in C_s}\max_{a\in A}(|r_{s'}^a - r_s^a| + \gamma|\sum_{D\in S'}\sum_{u\in D} P_{s'u}^a V_i(D) - \sum_{u\in S} P_{su}^a V_i(u)|) \\
&\le \frac{1}{|C_s|}\sum_{s'\in C_s}\max_{a\in A}(|r_{s'}^a - r_s^a| + \gamma|\sum_{u\in S} P_{s'u}^a V_i(C_u) - P_{su}^a V_i(u)|) \\
&\le \frac{1}{|C_s|}\sum_{s'\in C_s}\max_{a\in A}(|r_{s'}^a - r_s^a| + \gamma|\sum_{u\in S}(P_{s'u}^a - P_{su}^a)V_i(u)| + \gamma|\sum_{u\in S} P_{s'u}^a(V_i(C_u) - V_i(u))|) \\
&\le \frac{1}{c_R|C_s|}\sum_{s'\in C_s}\max_{a\in A}(c_R d_a(s,s') + c_T|\sum_{u\in S}(P_{s'u}^a - P_{su}^a)(\frac{c_R\gamma}{c_T}V_i(u))|) \\
&\qquad + \frac{\gamma}{|C_s|}\sum_{s'\in C_s}\max_{a\in A}\sum_{u\in S} P_{s'u}^a|V_i(C_u) - V_i(u)|
\end{aligned}
$$

Notice that by theorem 3.2.1 $\{\frac{c_R\gamma}{c_T}V_i(u) : u \in S\}$ constitutes a feasible solution for the primal LP for $m_i(P^a_s, P^a_{s'})$. Hence, we may continue by noting that the preceding expression is bounded above by:

$$
\begin{aligned}
&\leq \frac{1}{c_R|C_s|}\sum_{s'\in C_s}\max_{a\in A}(c_R d_a(s,s') + c_T m_i(P^a_s, P^a_{s'})) \\
&\qquad + \frac{\gamma}{|C_s|}\sum_{s'\in C_s}\max_{a\in A}\sum_{u\in S}P_{s'u}\max_{u\in S}|V_i(C_u) - V_i(u)| \\
&\leq \frac{1}{c_R|C_s|}\sum_{s'\in C_s} m_{i+1}(s,s') + \gamma\max_{u\in S}|V_i(C_u) - V_i(u)| \\
&\leq \frac{avg(s, m_{i+1})}{c_R} + \gamma\max_{u\in S}(\frac{1}{c_R}(avg(u,m_i) + \sum_{k=1}^{i-1}\gamma^{i-k}\max_{v\in S} avg(v,m_k))) \\
&\leq \frac{1}{c_R}(avg(s,m_{i+1}) + \gamma\max_{u\in S} avg(u,m_i) + \sum_{k=1}^{i-1}\gamma^{i+1-k}\max_{u\in S} avg(u,m_k)) \\
&\leq \frac{1}{c_R}(avg(s,m_{i+1}) + \sum_{k=1}^{i}\gamma^{(i+1)-k}\max_{u\in S} avg(u,m_k))
\end{aligned}
$$

□

Consider the use of clustering as an aggregation method. Roughly speaking, we choose certain seed states and for each, consider the class of states within distance $\epsilon$ for some some fixed positive $\epsilon$ (while ensuring that each state is placed in only one class). Then for a cluster $C$ and any state $s$ belonging to it, the above tells us that $|V^*(C) - V^*(s)| \leq \frac{2\epsilon}{c_R(1-\gamma)}$, provided $\gamma \leq c_T$. Thus, as $\epsilon$ decreases, the optimal values of a class and its states converge.

### 3.2.3 Restricting the Policy Space

The bounds above relate the distance assigned to states by a bisimulation metric to the values of those states under an optimal policy. What can we say when dealing with an arbitrary randomized stationary policy? In general, such a policy could dictate vastly different strategies for states which are bisimilar, and so we should expect no useful information in these cases. This suggests restricting our attention to those policies whose behaviour is governed by the distances assigned by bisimulation metrics. Formally, we have the following result:

**Theorem 3.2.3.** *Let $\pi \in \Pi^{RS}$ such that for some $c_\pi \geq 0$,*

$$\max_{a \in A} |\pi(s,a) - \pi(s',a)| \leq c_\pi m(s,s') \quad \forall s,\, s' \in S$$

*Let $\gamma \leq c_T$. Then*

1. $|V_i^\pi(s) - V_i^\pi(s')| \leq x_i m_i(s,s') + y_i$, $\forall i \geq 0$.

2. $|V^\pi(s) - V^\pi(s')| \leq x m(s,s')$

*where*

$$y_0 = 0,\; y_i = \gamma y_{i-1} + \frac{|A| c_\pi (1-\gamma^i) c_T^{i-1}}{(1-\gamma)}$$

*and*

- *if $c_\pi = 0$ then $x_i = \frac{1}{c_R}$ for $i \geq 0$ and $x = \frac{1}{c_R}$*

- *if $0 < c_\pi \leq \frac{(1-\gamma)(c_T-\gamma)}{\gamma |A| c_R}$ and $\gamma < c_T$ then $x_i = \frac{1}{c_R} + \frac{|A| c_\pi (1-\gamma^i)}{(1-\gamma)}$ for $i \geq 0$ and $x = \frac{1}{c_R} + \frac{|A| c_\pi}{(1-\gamma)}$*

- *if* $0 < \frac{(c_T-\gamma)}{\gamma|A|c_R} < c_\pi$ *and* $\gamma < c_T$ *then* $x_0 = \frac{c_T}{\gamma c_R}$, $x_i = \frac{\gamma}{c_T}x_{i-1} + \frac{|A|c_\pi(1-\gamma^i)}{(1-\gamma)}$ *for* $i \geq 1$, *and* $x = \frac{|A|c_\pi c_T}{(1-\gamma)(c_T-\gamma)}$

*Proof.* The second item will follow from the first by taking limits. To establish the first we proceed by induction. Before we do so, note that by induction $m(s,s') - m_i(s,s') \leq c_T^i$ for $i \geq 0$, so that for every $s, s' \in S$,

$$\max_{a\in A} |\pi(s,a) - \pi(s',a)| \leq c_\pi m(s,s') \leq c_\pi(m_i(s,s') + c_T^i)$$

Now the base case holds vacuously as for any $s, s' \in S$, $V_0^\pi(s)$, $m_0(s,s')$, and $y_0$ are all zero. For the inductive step, note:

$$\begin{aligned}
&|V_{i+1}^\pi(s) - V_{i+1}^\pi(s')| \\
&= |\sum_{a\in A}\pi(s,a)(r_s^a + \gamma\sum_{u\in S}P_{su}^a V_i^\pi(u)) - \sum_{a\in A}\pi(s',a)(r_{s'}^a + \gamma\sum_{u\in S}P_{s'u}^a V_i^\pi(u))| \\
&= |\sum_{a\in A}\pi(s,a)(r_s^a - r_{s'}^a + \gamma\sum_{u\in S}(P_{su}^a - P_{s'u}^a)V_i^\pi(u)) \\
&\qquad\qquad + \sum_{a\in A}(\pi(s,a) - \pi(s',a))(r_{s'}^a + \gamma\sum_{u\in S}P_{s'u}^a V_i(u))| \\
&\leq \max_{a\in A}|r_s^a - r_{s'}^a + \gamma\sum_{u\in S}(P_{su}^a - P_{s'u}^a)V_i^\pi(u))| + |A|\max_{a\in A}|\pi(s,a) - \pi(s',a)|\frac{(1-\gamma^{i+1})}{(1-\gamma)} \\
&\leq \max_{a\in A}(d_a(s,s') + \gamma|\sum_{u\in S}(P_{su}^a - P_{s'u}^a)V_i^\pi(u)|) + |A|c_\pi m(s,s')\frac{(1-\gamma^{i+1})}{(1-\gamma)}
\end{aligned}$$

By the induction hypothesis we have

$$\frac{(1-\gamma^i)}{(1-\gamma)}|V_i^\pi(s) - V_i^\pi(s')| \leq \frac{(1-\gamma^i)}{(1-\gamma)}x_i m_i(s,s') + \frac{(1-\gamma^i)}{(1-\gamma)}y_i$$

and $0 \leq V^\pi(s) \leq 1$ so that $\{\frac{(1-\gamma^i)}{(1-\gamma)}V_i^\pi(s) | s \in S\}$ constitutes a feasible solution to the primal Kantorovich LP with cost function $\frac{(1-\gamma^i)}{(1-\gamma)}x_i m_i(s,s') + \frac{(1-\gamma^i)}{(1-\gamma)}y_i$.

Thus, by lemma 2.4.5 the preceding quantity is bounded above by:

$$
\begin{aligned}
&\leq \max_{a \in A}(d_a(s,s') + \gamma(x_i m_i(P_s^a, P_{s'}^a) + y_i)) + |A| c_\pi m(s,s') \frac{(1-\gamma^{i+1})}{(1-\gamma)} \\
&\leq \frac{m_{i+1}(s,s')}{c_R} + \max_{a \in A}((\frac{\gamma}{c_T} x_i - \frac{1}{c_R}) c_T m_{i+1}(P_s^a, P_{s'}^a)) \\
&\qquad\qquad + \gamma y_i + |A| c_\pi (m_{i+1}(s,s') + c_T^{i+1}) \frac{(1-\gamma^{i+1})}{(1-\gamma)} \\
&\leq x_{i+1} m_{i+1}(s,s') + y_{i+1}
\end{aligned}
$$

where

$$y_{i+1} = \gamma y_i + \frac{|A| c_\pi (1-\gamma^{i+1}) c_T^{i+1}}{(1-\gamma)}$$

and

$$x_{i+1} = \begin{cases} \frac{1}{c_R} + \frac{|A| c_\pi (1-\gamma^{i+1})}{(1-\gamma)} & \text{if } x_i < \frac{c_t}{\gamma c_R} \\ \frac{\gamma}{c_T} x_i + \frac{|A| c_\pi (1-\gamma^{i+1})}{(1-\gamma)} & \text{otherwise} \end{cases}$$

By induction, we may rewrite

$$y_{i+1} = \frac{|A| c_\pi c_T^{i+1}}{(1-\gamma)} \{ \frac{1-(\frac{\gamma}{c_T})^{i+1}}{1-\frac{\gamma}{c_T}} - \frac{\gamma^{i+1} - (\frac{\gamma}{c_T})^{i+1}}{1-\frac{1}{c_T}} \}$$

so that $\lim_{i \to \infty} y_i = 0$.

- If $c_\pi = 0$ then $x_0 = \frac{1}{c_R} \leq \frac{c_t}{\gamma c_R}$ so that $x_i = \frac{1}{c_R}$ for every $i$ and $x = \lim_{i \to \infty} x_i = \frac{1}{c_R}$.
- If $0 < c_\pi \leq \frac{(1-\gamma)(c_T-\gamma)}{\gamma |A| c_R}$ and $\gamma < c_T$ then

$$
\begin{aligned}
c_\pi \leq \frac{(1-\gamma)(c_T-\gamma)}{\gamma |A| c_R} \quad &\Rightarrow \quad 1 \leq \frac{(1-\gamma)(c_T-\gamma)}{\gamma |A| c_R c_\pi} \\
&\Rightarrow \quad 1 - \gamma^i < \frac{(1-\gamma)(c_T-\gamma)}{\gamma |A| c_R c_\pi} \text{ for every } i \\
&\Rightarrow \quad \frac{1}{c_R} + \frac{|A| c_\pi (1-\gamma^i)}{(1-\gamma)} < \frac{c_T}{\gamma c_R} \text{ for every } i
\end{aligned}
$$

Since $x_0 = \frac{1}{c_R} < \frac{c_T}{\gamma c_R}$ it follows by induction that $x_i = \frac{1}{c_R} + \frac{|A|c_\pi(1-\gamma^i)}{(1-\gamma)}$ for $i \geq 0$ and $x = \lim_{i\to\infty} x_i = \frac{1}{c_R} + \frac{|A|c_\pi}{(1-\gamma)}$.

- if $0 < \frac{(c_T-\gamma)}{\gamma|A|c_R} < c_\pi$ and $\gamma < c_T$ then since $x_0 = \frac{c_T}{\gamma c_R}$ it follows by induction that $x_i \geq \frac{c_T}{\gamma c_R}$ for every $i$ and $x_i = \frac{\gamma}{c_T}x_{i-1} + \frac{|A|c_\pi(1-\gamma^i)}{(1-\gamma)}$ for $i \geq 1$. In fact, using induction once more yields

$$x_{i+1} = (\frac{\gamma}{c_T})^{i+1}x_0 + \frac{|A|c_\pi}{1-\gamma}\{\frac{1-(\frac{\gamma}{c_T})^{i+1}}{1-\frac{\gamma}{c_T}} - \frac{\gamma^{i+1}-(\frac{\gamma}{c_T})^{i+1}}{1-\frac{1}{c_T}}\}$$

so that $x = \lim_{i\to\infty} x_i = \frac{|A|c_\pi c_T}{(1-\gamma)(c_T-\gamma)}$.

□

Naturally, we can extend such bounds to an aggregate MDP $M'$. Extend $\pi$ to $\pi'$ on $M'$ by averaging over the states in an equivalence class, i.e.

$$\pi'(C,a) = \frac{1}{|C|}\sum_{s\in C}\pi(s,a).$$

Then:

**Theorem 3.2.4.** *Suppose $\gamma \leq c_T$. Let $\{x_i\}$, $\{y_i\}$, and $x$ be as in the previous theorem. Then $\forall s \in S$:*

1. $\forall i \geq 0$,

$$|V_i^{\pi'}(C_s) - V_i^{\pi}(s)| \leq x_i avg(s,m_i) + y_i + \sum_{k=1}^{i-1}\gamma^{i-k}(x_k \max_{u\in S} avg(u,m_k) + y_k)$$

2. $|V^{\pi'}(C_s) - V^{\pi}(s)| \leq x(avg(s,m) + \frac{\gamma}{1-\gamma}\max_{u\in S} avg(u,m))$

This follows by emulating the proof of theorem 3.2.2 with use of the bounds in the preceding theorem.

Of course the particular situation of interest lies in using a policy for the aggregate MDP $M'$ to recover one for the original MDP $M$ along with a bound on the associated values of each. By working backwards and using the results proven above we can do just this. Assume that the partition of the original MDP satisfies the requirement that if two states are bisimilar then they belong to the same equivalence class. For example, such is always the case when we cluster using a bisimulation metric. Given $\pi'$ for $M'$, define $\pi$ for $M$ by

$$\pi(s,a) = \pi'(C_s, a).$$

Then it follows that for a given bisimulation metric $m$,

$$\max_{a \in A} |\pi(s,a) - \pi(s',a)| \leq bis(s,s') \leq c_\pi m(s,s) \quad \forall s,\ s' \in S$$

where the latter inequality follows from the equivalence of all bisimulation metrics. Moreover,

$$\pi'(C,a) = \frac{1}{|C|} \sum_{s \in C} \pi'(C,a) = \frac{1}{|C|} \sum_{s \in C} \pi(s,a)$$

so that the bounds proven above hold here.

## 3.3 An Algorithm

If our family of RL metrics and associated bounds are to be of any use then we must have some way of computing the metrics. Our method of constructing the bisimulation metrics has been carried out with this in mind. The idea is that for a given RL metric $m$ we can approximate its distances via the iterates $\{m_i\}$. By induction we find that for every $i$, $m(s,s') - m_i(s,s') \leq c_T^i$. Thus, to calculate distances up to a prescribed degree of accuracy $\delta$ we need only iterate for $i = \lceil \frac{\ln \delta}{\ln c_T} \rceil$ steps.

The pseudocode in figure 3.3 shows how to calculate the distances to within error $\delta$. Given that each Kantorovich subproblem can be solved in time $O(|S|^2 \log |S|)$ the entire computation has running time $O(|A||S|^4 \log |S| \frac{\ln \delta}{\ln c_T})$.

```
(INITIALIZATION)
For $s, s' = 1$ to $N$ do
    $m(s,s') = 0$
    For $a = 1$ to $|A|$ do
        $d_a(s,s') = abs(r_s^a - r_{s'}^a)$
(MAIN LOOP)
For $i = 1$ to $\lceil \frac{\ln \delta}{\ln c_T} \rceil$ do
    For $s, s' = 1$ to $N$ do
        For $a = 1$ to $A$ do
            $ProbDist_a(s,s') = \min_{l_{kj}} \sum_{k,j} l_{kj} m(k,j)$
                        subject to:  $\sum_j l_{kj} = P_{sk}^a$
                                     $\sum_k l_{kj} = P_{s'j}^a$
                                     $l_{kj} \geq 0$
    For $s, s' = 1$ to $N$ do
        $m(s,s') = \max_a (c_R d_a(s,s') + c_T ProbDist_a(s,s'))$
```

Figure 3.3: Pseudocode to compute RL metric distances to within error $\delta$.

# Chapter 4

# Experimental Results

In this chapter we investigate RL metrics in practice. Specifically, our toy experiments demonstrate the accuracy of value function bounds on MDPs and associated aggregate MDPs obtained by clustering with RL metric distances.

## 4.1 Random MDPs

The implementation of random finite MDPs used here is based upon the design advocated by Sutton and Kautz in [13]. A random MDP is generated by specifying the size of an enumerated state space, $S$, the size of an enumerated action space, $A$, and a branching factor $B$. The branching factor specifies the number of possible next states for each state. These next states are chosen uniformly random from the $|S|$ states. The $B$ transition probabilities are chosen as a random partition of $[0, 1]$. The expected immediate reward for each state and action is then chosen according to a normal distribution with mean 0 and variance 1. For convenience we then shift and scale these rewards into $[0, 1]$.

## 4.2 Software

Distance calculating code was primarily implemented in Java. The minimum cost flow solver used was MCFZIB, a network simplex algorithm implementation by Andreas Löbel. This particular C++ implementation is part of the MCFClass Project found online at `http://www.di.unipi.it/di/groups/optimize/Software/MCF.html`.

## 4.3 Experiments and Results

Our experiments are designed to demonstrate the usefulness of the RL metrics in the case where RL metric distances are explicitly used for state space aggregation. We do so by examining two quantities: tightness of the optimal value bound for aggregation as given in the previous chapter, and the variation in size of the aggregate MDP. Let us elaborate: we fix a setting of parameters $|S| = 25$ (size of state space), $|A| \in \{2, 5, 10\}$ (size of action space), $B \in \{2, 5, 10\}$ (branching factor), and $\gamma \in \{0.1, 0.5, 0.9\}$ (discount factor). For each parameter setting, 100 random MDPs are generated using the procedure described above. For each MDP we compute the RL metric distances with $c_T = \gamma$ and $c_R = 1 - \gamma$. Next we fix a cluster size, $\epsilon$, chosen from $[0, 1]$ in steps of 0.01. We obtain an aggregate MDP for this $\epsilon$ by choosing the first state of the enumerated state space and clustering together all states within RL distance $\epsilon$. We then take the next state not already within a cluster and repeat the clustering until the entire state space has been partitioned. Finally, we assign rewards and transition probabilities by averaging these values over clusters, as described in the previous chapter. [1] We now compute the optimal values of the states in the original MDP and in the aggregate MDP using the traditional dynamic programming method. This allows us to compare the *exact* difference between the optimal values of a state and its equivalence class (cluster) with the theoretical bound provided in the previous chapter. Specifically, we look at

$$\max_{s \in S} \frac{1}{c_R}(avg(s,m) + \frac{\gamma}{1-\gamma} \max_{u \in S} avg(u,m)) - |V^*(C_s) - V^*(s)|$$

where $c_T = \gamma$, $c_R = 1 - \gamma$, $m$ is the corresponding RL metric, and $C_s$ is the equivalence class of the aggregate MDP containing state $s$. We also look at the size of the aggregate MDP. Both these quantities are averaged over the 100 random MDPs. We then plot the averaged quantities of bound tightness, and aggregate MDP size as the cluster size varies. All data was plotted using 95% confidence intervals.

The results can be seen in figures 4.1 to 4.18. It is worthwhile for the reader, as he/she examines the graphical data, to compare our results to the case in which exact bisimulation alone would be used for compression. Figures 4.1 to 4.9 each display plots of the cluster size for an aggregate MDP versus the maximum error bound accuracy for a given setting of $|A|$ and $B$, and all values of $\gamma$ (recall $|S| = 25$ for all experiments). Figures 4.10 to 4.18

[1] Of course, this is not the only way to obtain an aggregate MDP. One could alternatively choose random seed states and cluster around these, or place states in more than one cluster and use weighted averages for rewards and transition probabilities. However, for purposes of illustration our simple clustering method is sufficient.

similarly each contain plots of the cluster size versus the size of the aggregate MDP. Note that the variations of the bound tightness and the variations of the MDP size are each typified by figures 4.1 and 4.10, and so it will be helpful to observe these as we analyze the results.

Observe that convergence indeed occurs smoothly, i.e. as the cluster size varies the change in the maximum error bound and the change in the size of the aggregate MDP is not too large for each. Additionally, note that as the cluster size tends to zero, the aggregate model "tends" to the original model as expected. By contrast, using exact bisimulation both the error bound and the MDP size would be maximal for all cluster sizes save 0.

The graphs of both data quantities additionally demonstrate the strong dependence of the data on $c_T$ ($= \gamma$ here). Note that for high values of $c_T$ (close to 1) the rate of change of the error bound tightness with $\epsilon$ increases much more quickly than for low values. The same can be said for the rate of change of the MDP size. This is somewhat troublesome as high values of $c_T$ would be more useful for practical applications. Thus, for such applications greater care must be taken when choosing a cluster size. Yet, it is not immediately clear how this can be done for a given parameter setting.

Nevertheless, we are still confident in the potential usefulness of the RL metrics. The graphical representations show that each parameter setting admits a range (decreasing in length with increases in $c_T$) of cluster sizes in which RL metric aggregation is advantageous, i.e. the tradeoff between the error and the aggregate MDP size is not too great. Moreover, the greater structure usually present in real world models could potentially improve the performance results. Of course, this, along with proper determination of cluster size, is something that must be further explored.

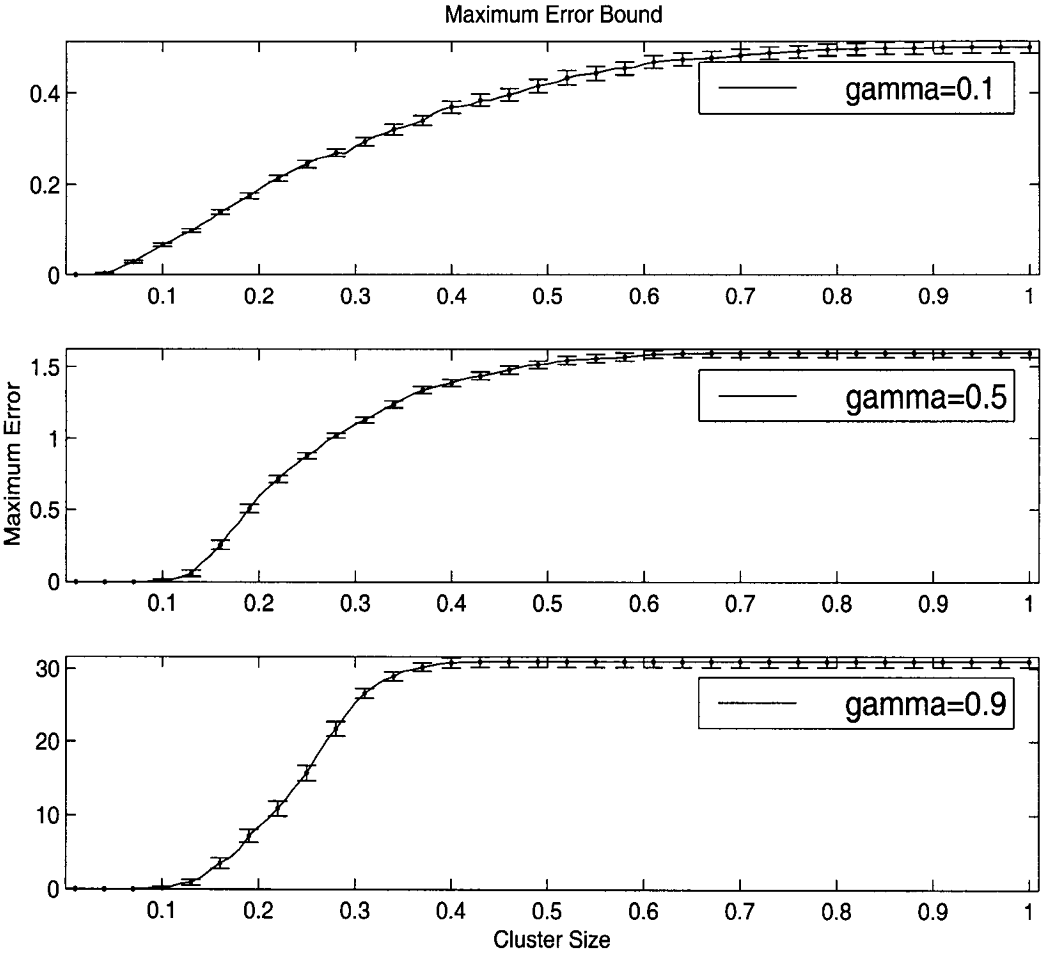

Figure 4.1: Error Bound Accuracy for $|S| = 25$, $|A| = 2$, $B = 2$

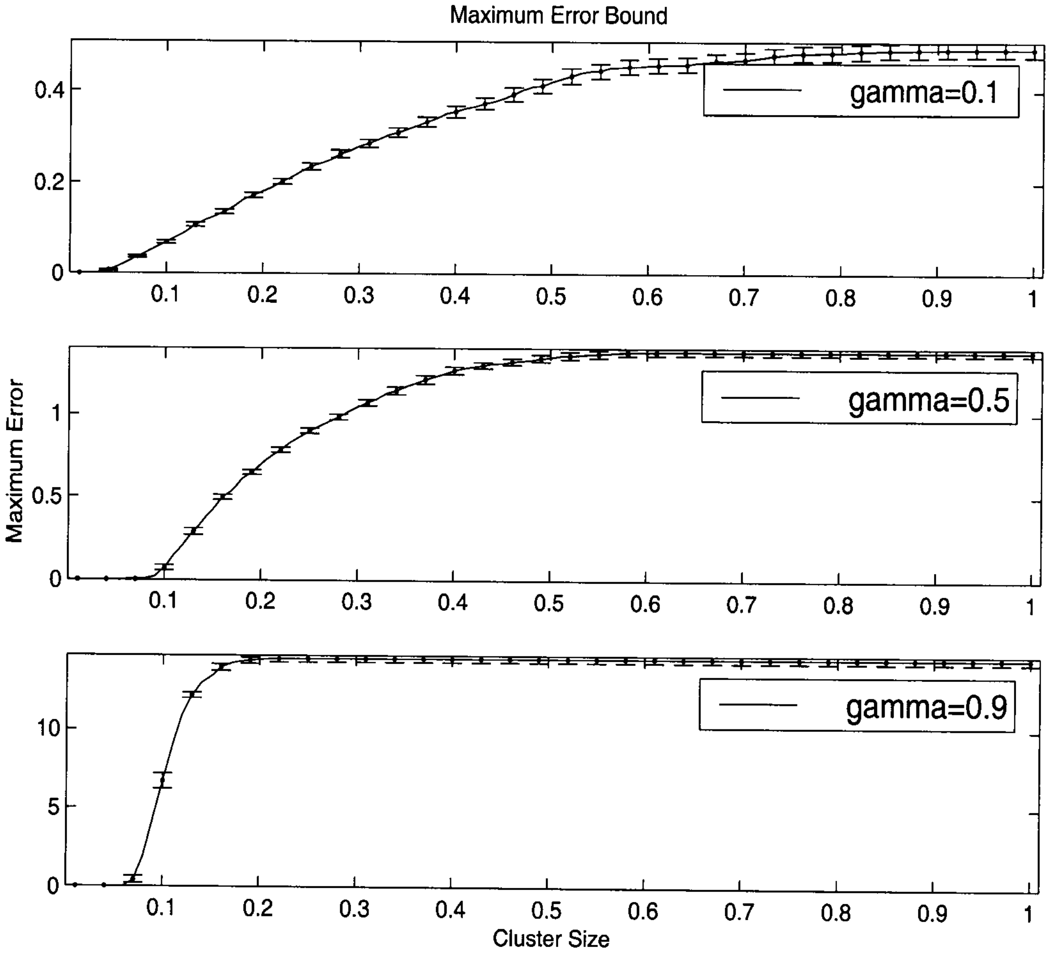

Figure 4.2: Error Bound Accuracy for $|S| = 25$, $|A| = 2$, $B = 5$

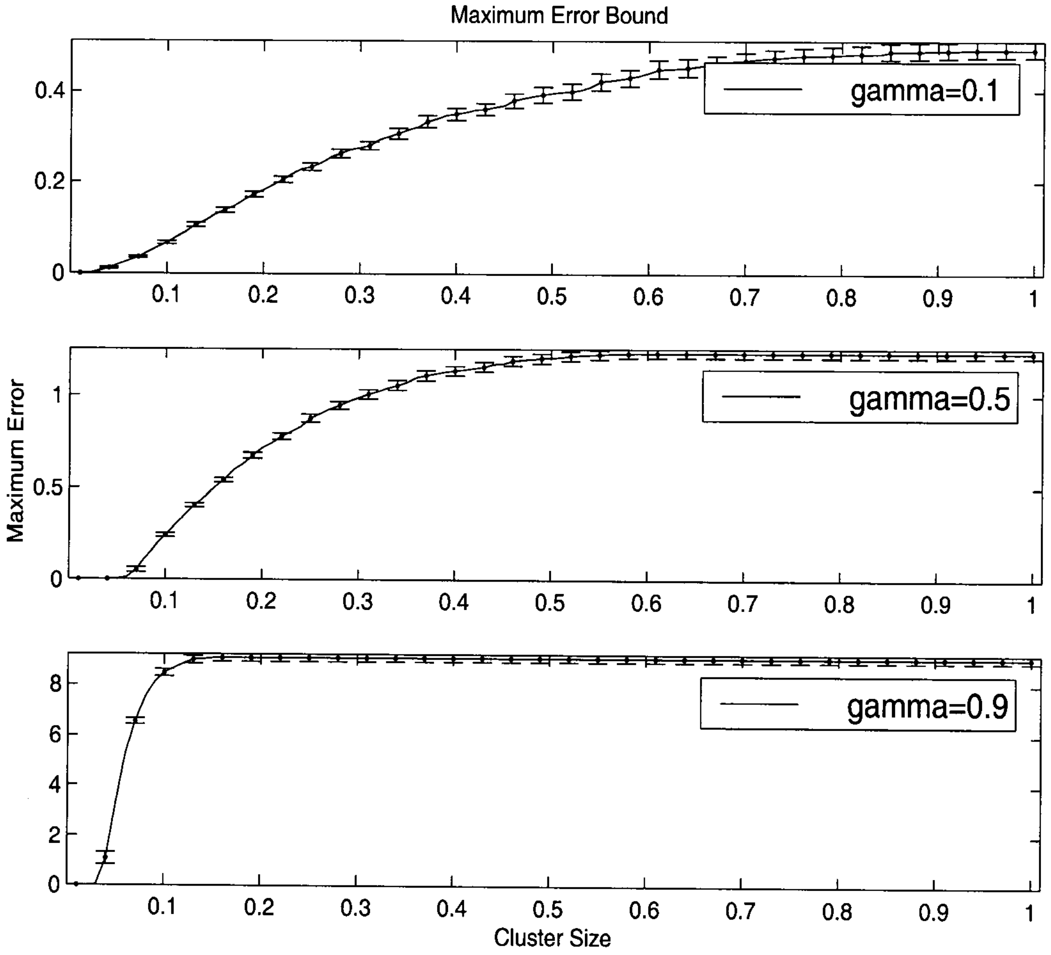

Figure 4.3: Error Bound Accuracy for $|S| = 25$, $|A| = 2$, $B = 10$

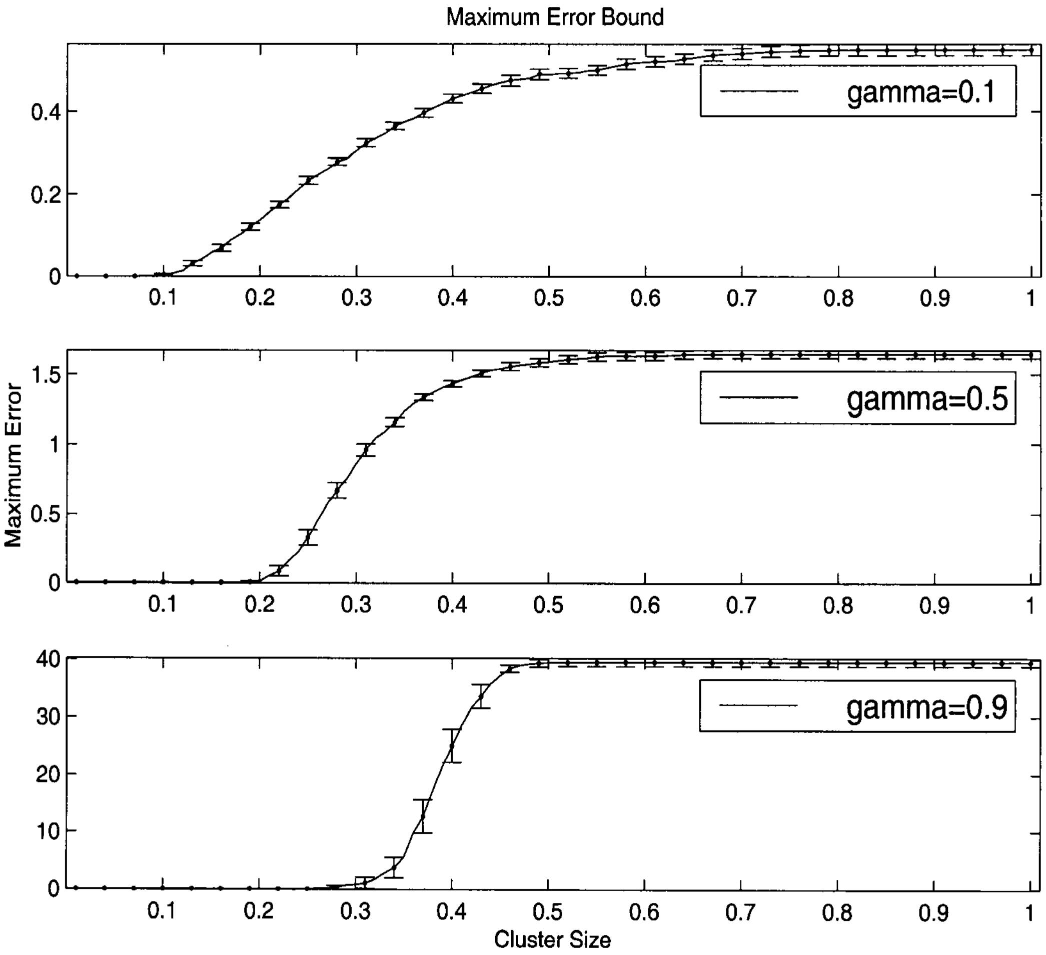

Figure 4.4: Error Bound Accuracy for $|S| = 25$, $|A| = 5$, $B = 2$

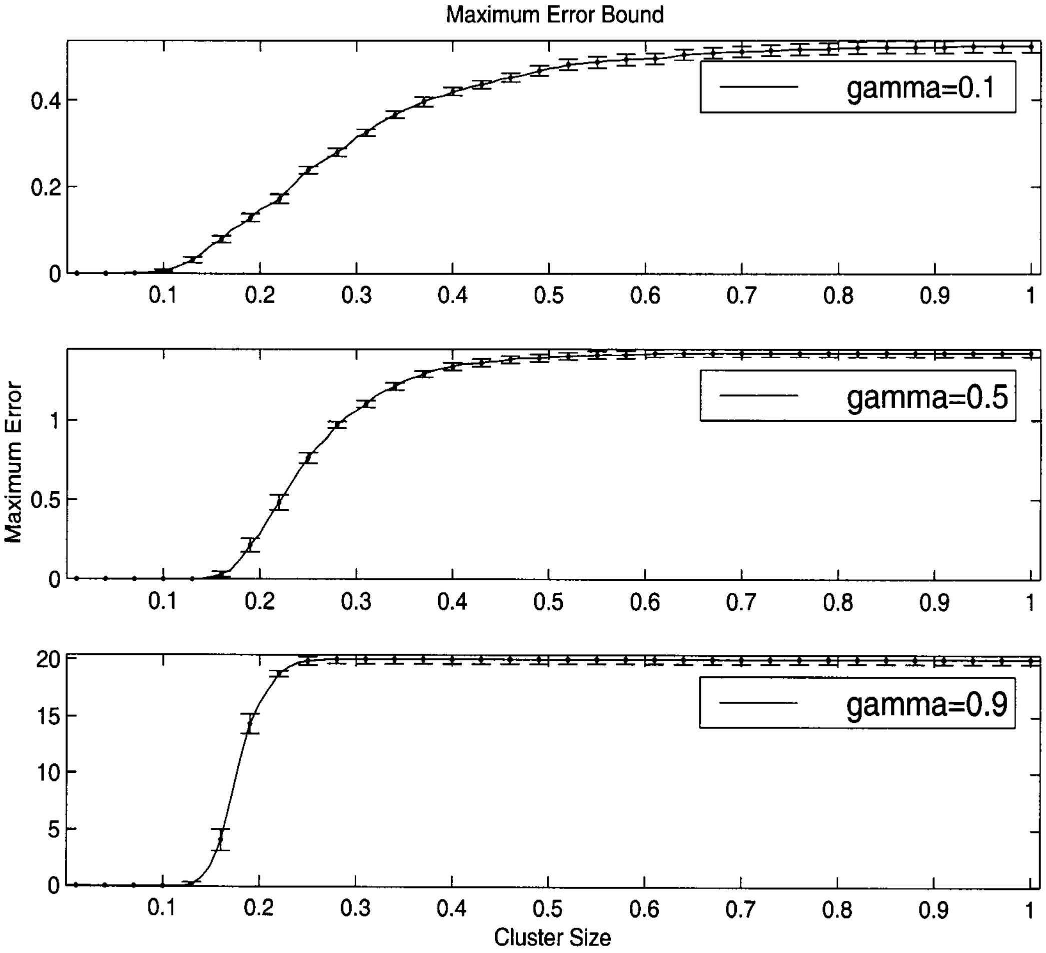

Figure 4.5: Error Bound Accuracy for $|S| = 25$, $|A| = 5$, $B = 5$

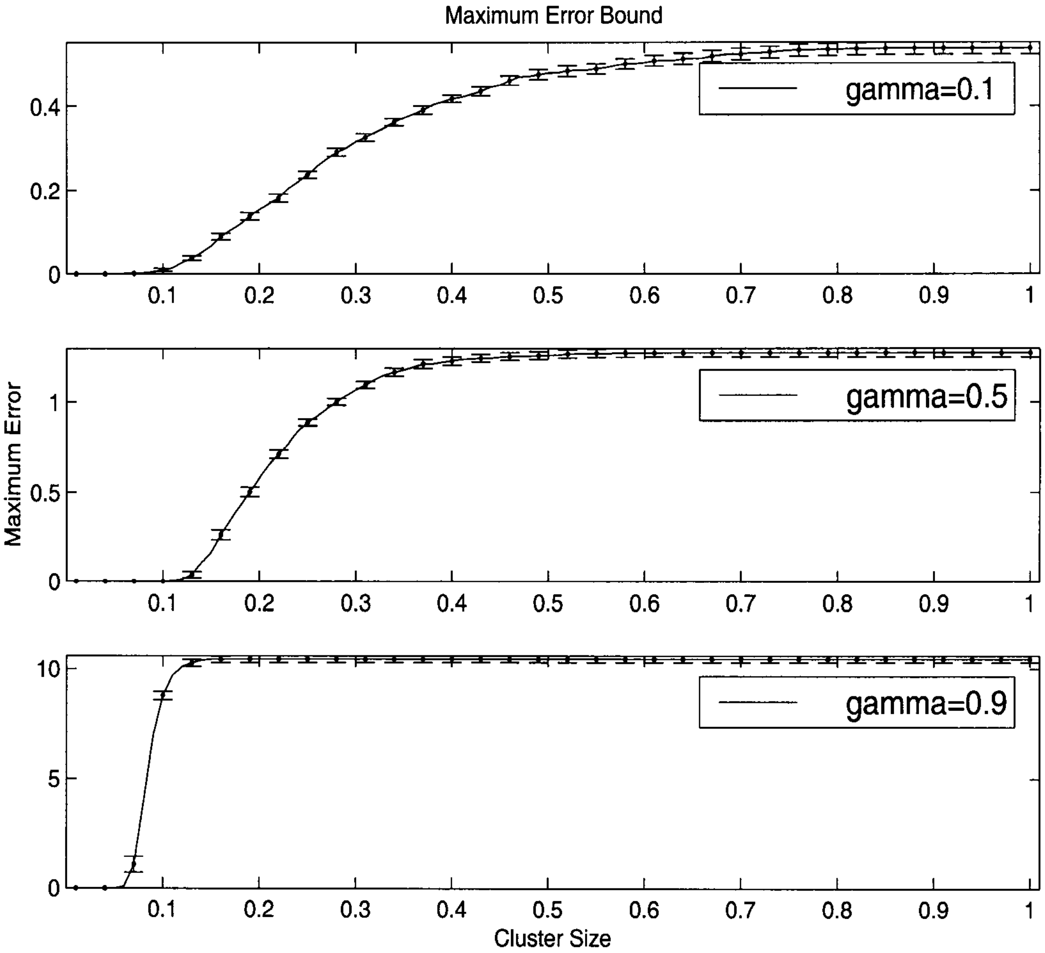

Figure 4.6: Error Bound Accuracy for $|S| = 25$, $|A| = 5$, $B = 10$

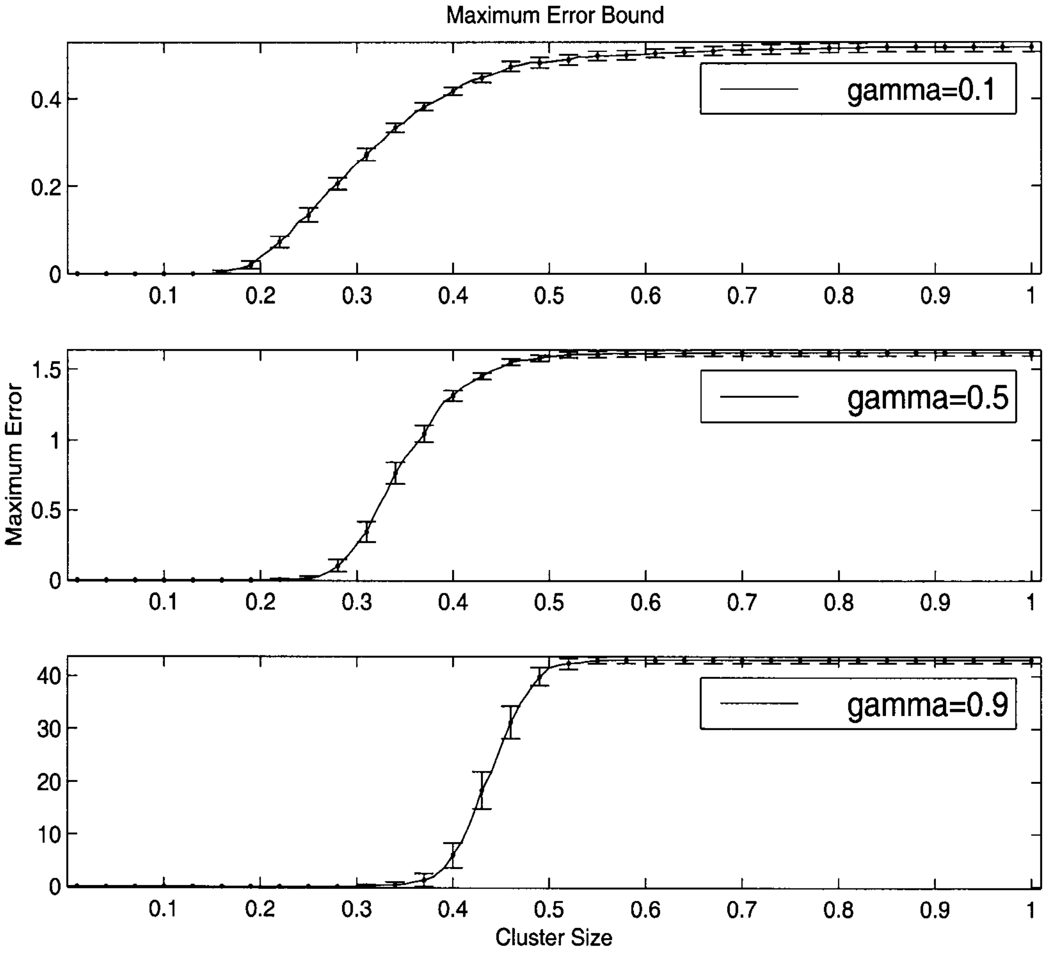

Figure 4.7: Error Bound Accuracy for $|S| = 25$, $|A| = 10$, $B = 2$

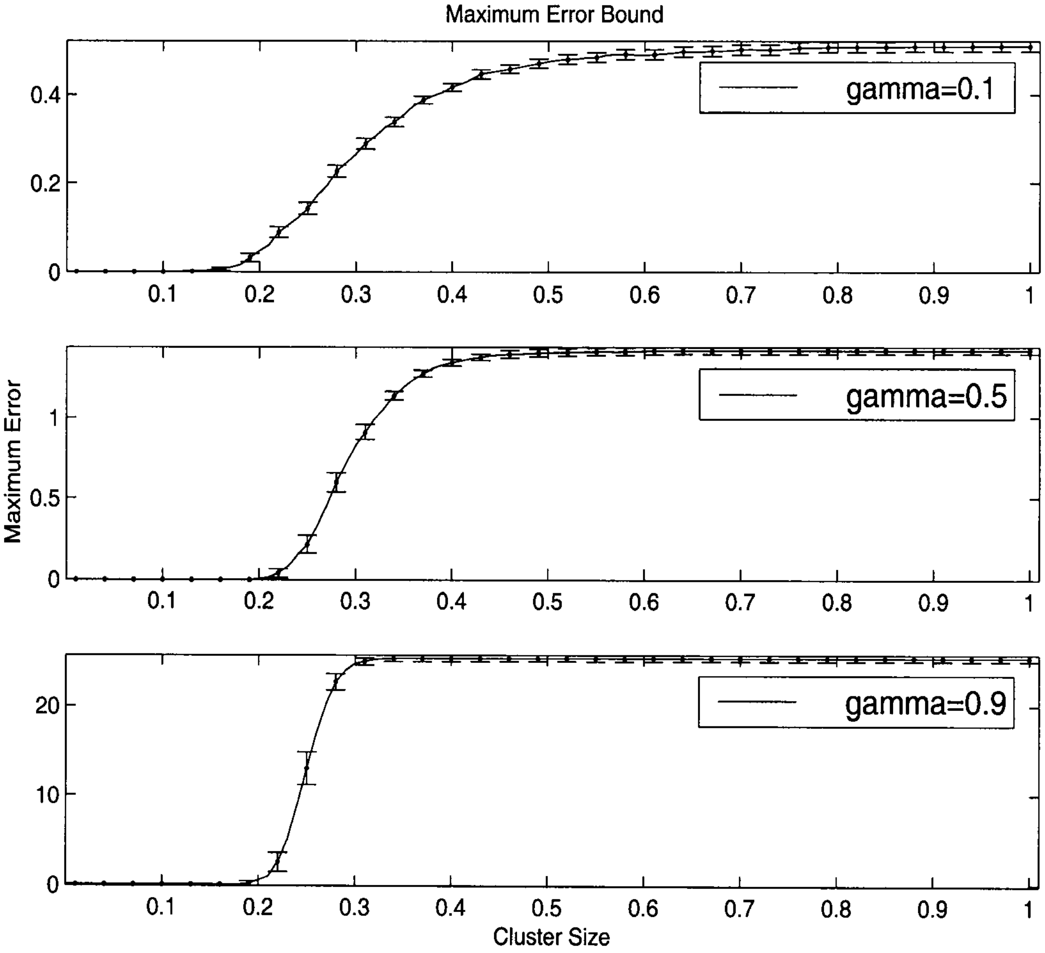

Figure 4.8: Error Bound Accuracy for $|S| = 25$, $|A| = 10$, $B = 5$

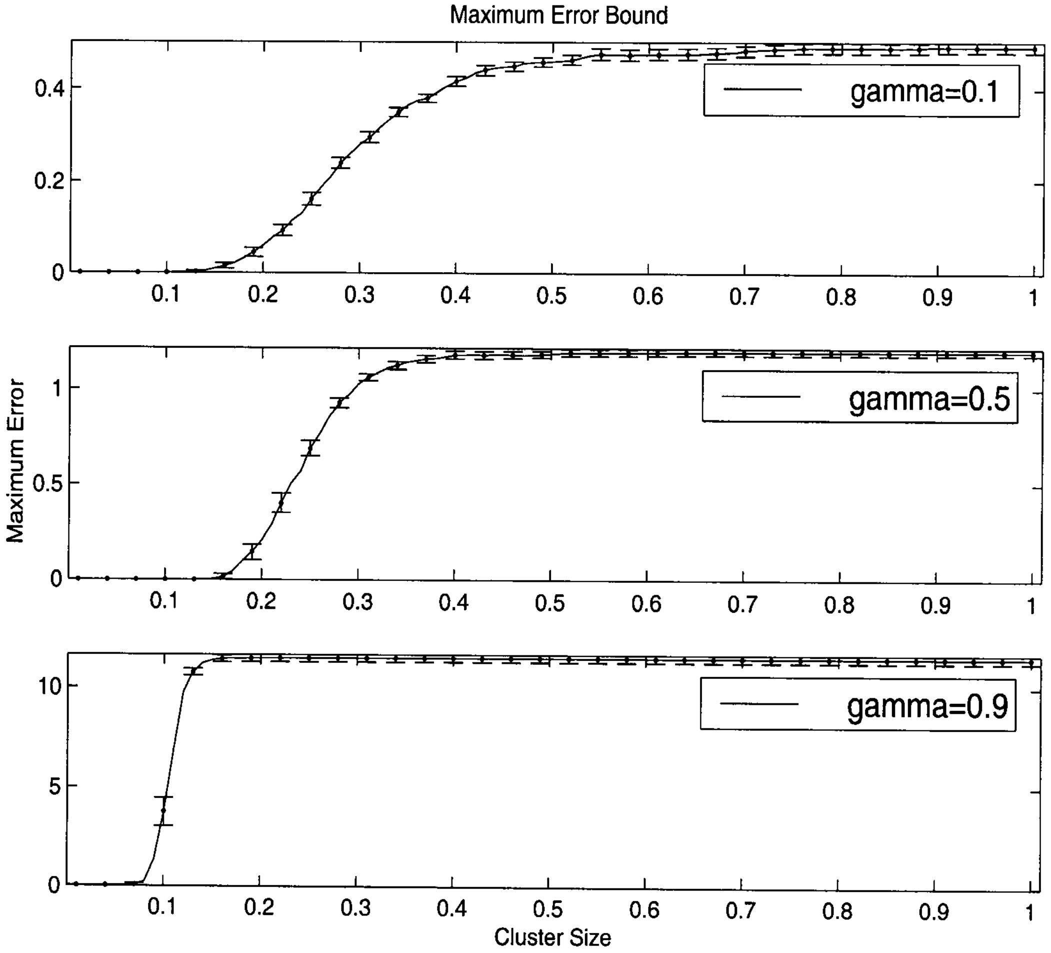

Figure 4.9: Error Bound Accuracy for $|S| = 25$, $|A| = 10$, $B = 10$

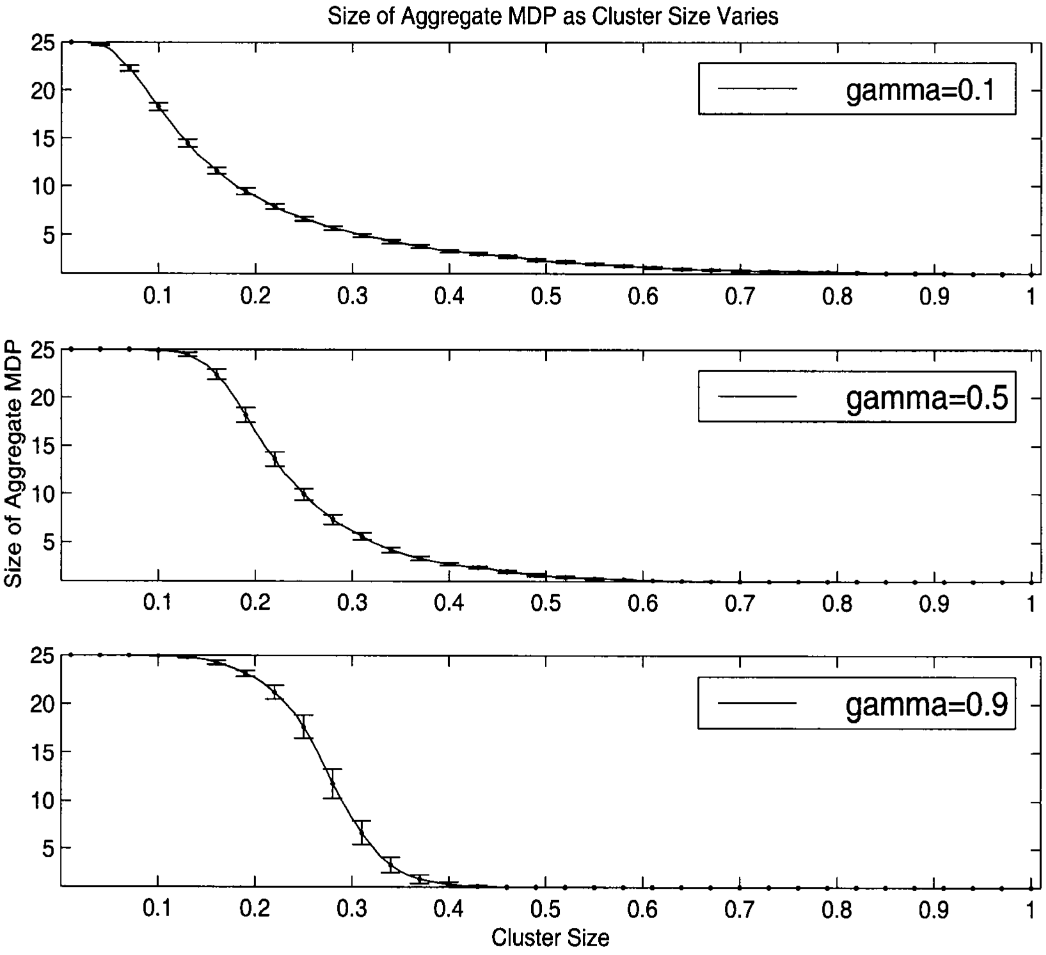

Figure 4.10: Aggregate MDP Sizes for $|S| = 25$, $|A| = 2$, $B = 2$

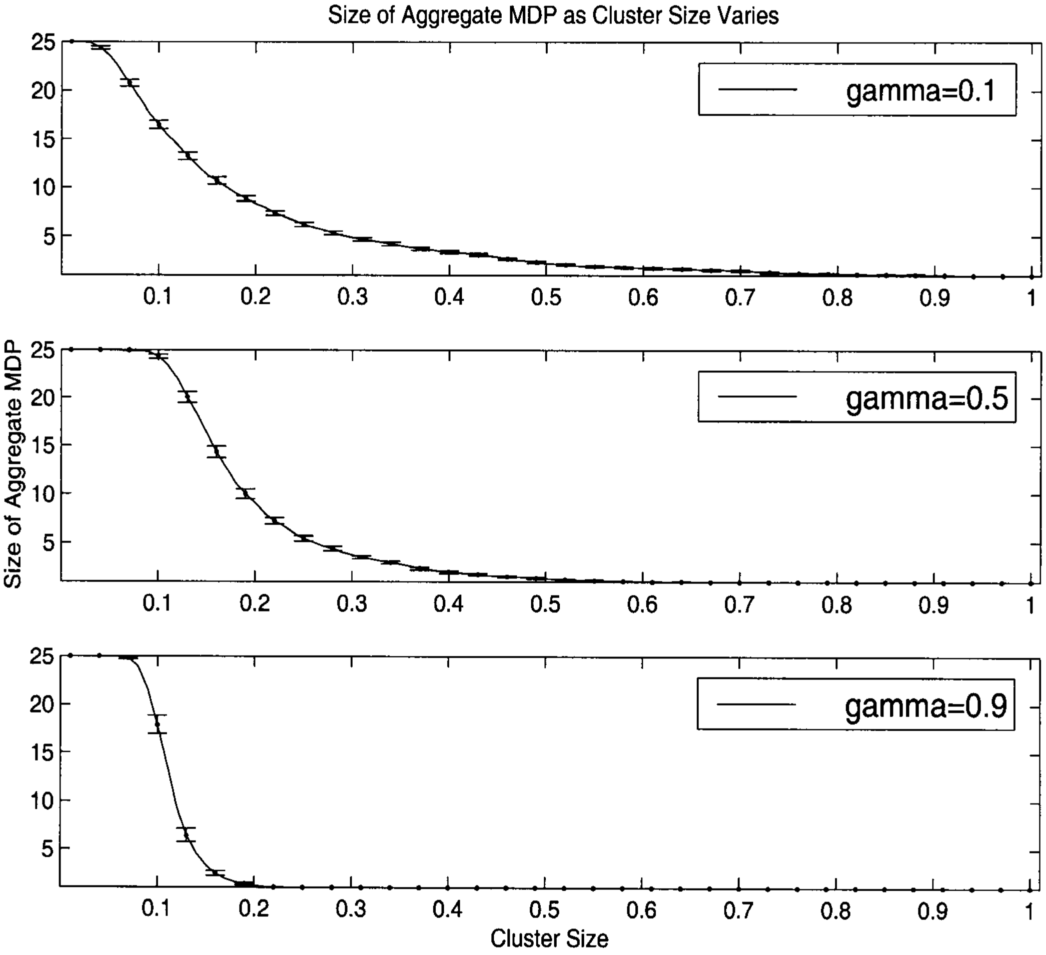

Figure 4.11: Aggregate MDP Sizes for $|S| = 25$, $|A| = 2$, $B = 5$

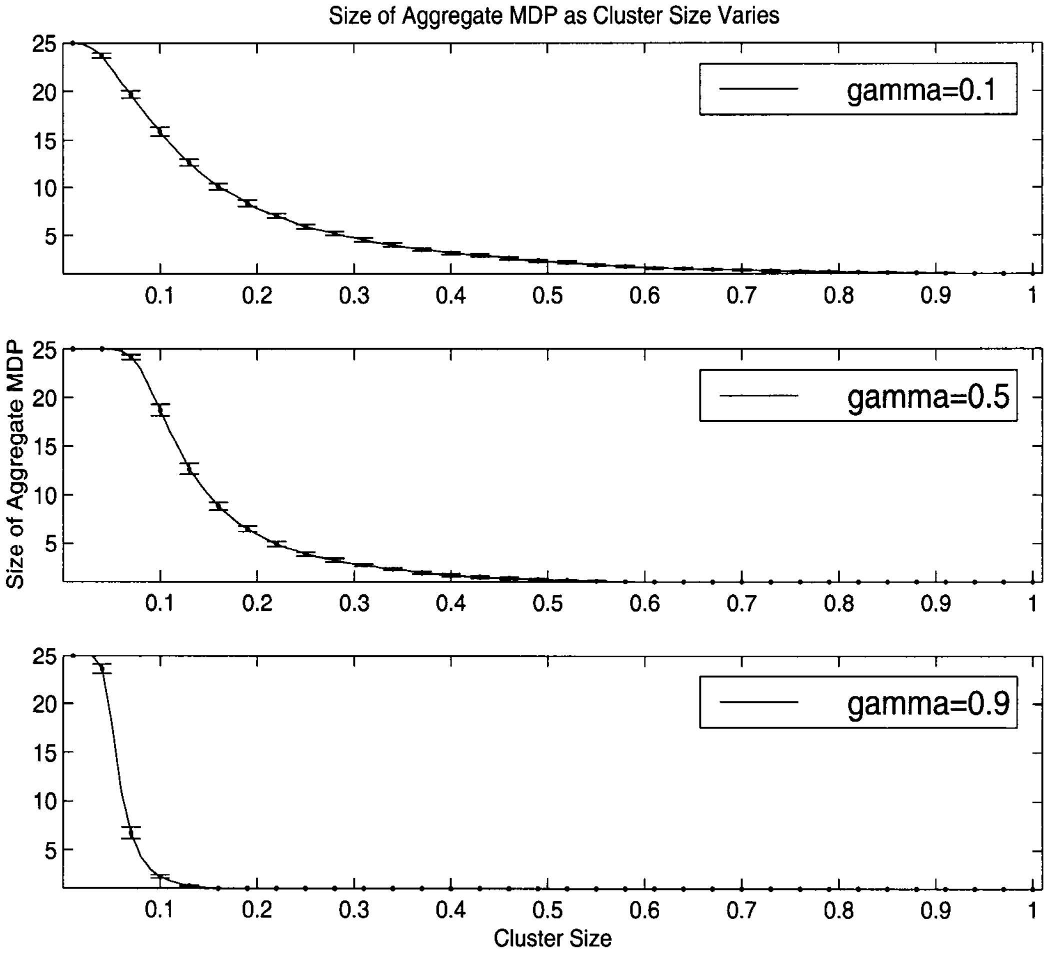

Figure 4.12: Aggregate MDP Sizes for $|S| = 25$, $|A| = 2$, $B = 10$

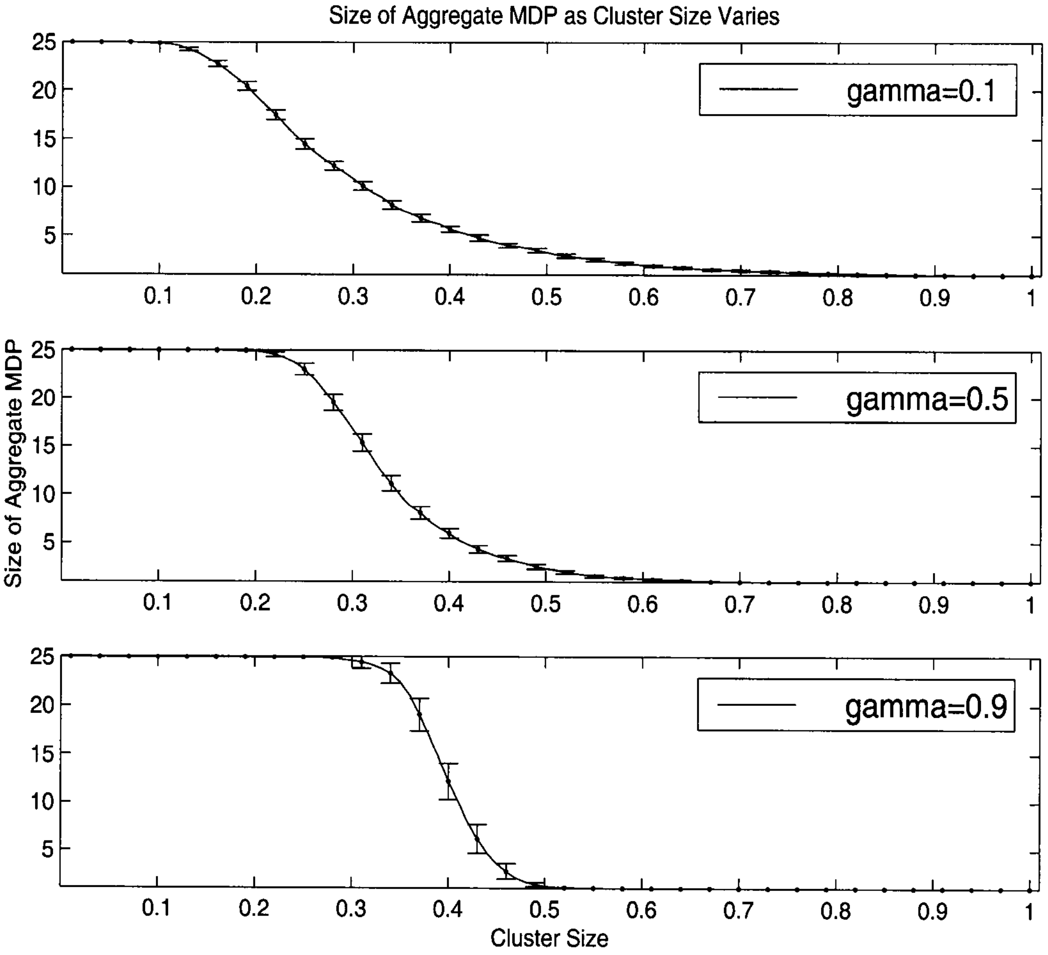

Figure 4.13: Aggregate MDP Sizes for $|S| = 25$, $|A| = 5$, $B = 2$

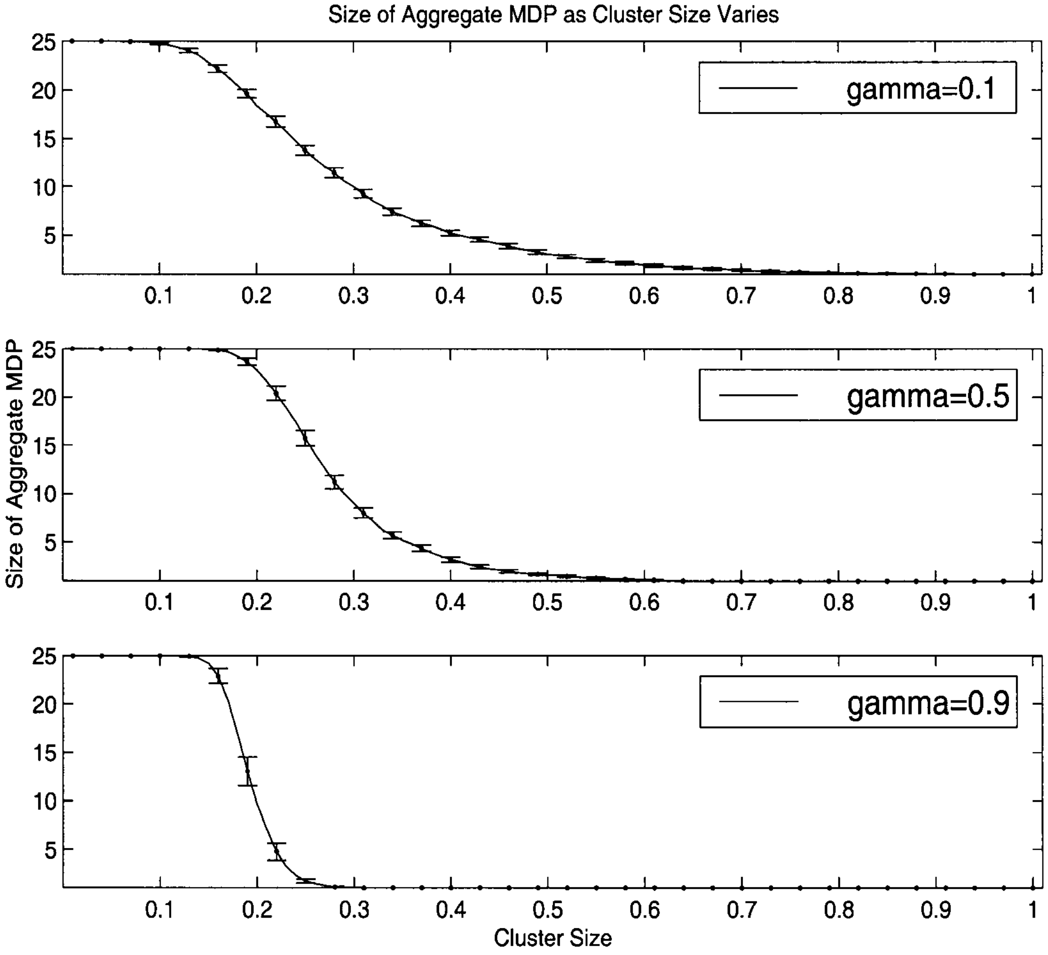

Figure 4.14: Aggregate MDP Sizes for $|S| = 25$, $|A| = 5$, $B = 5$

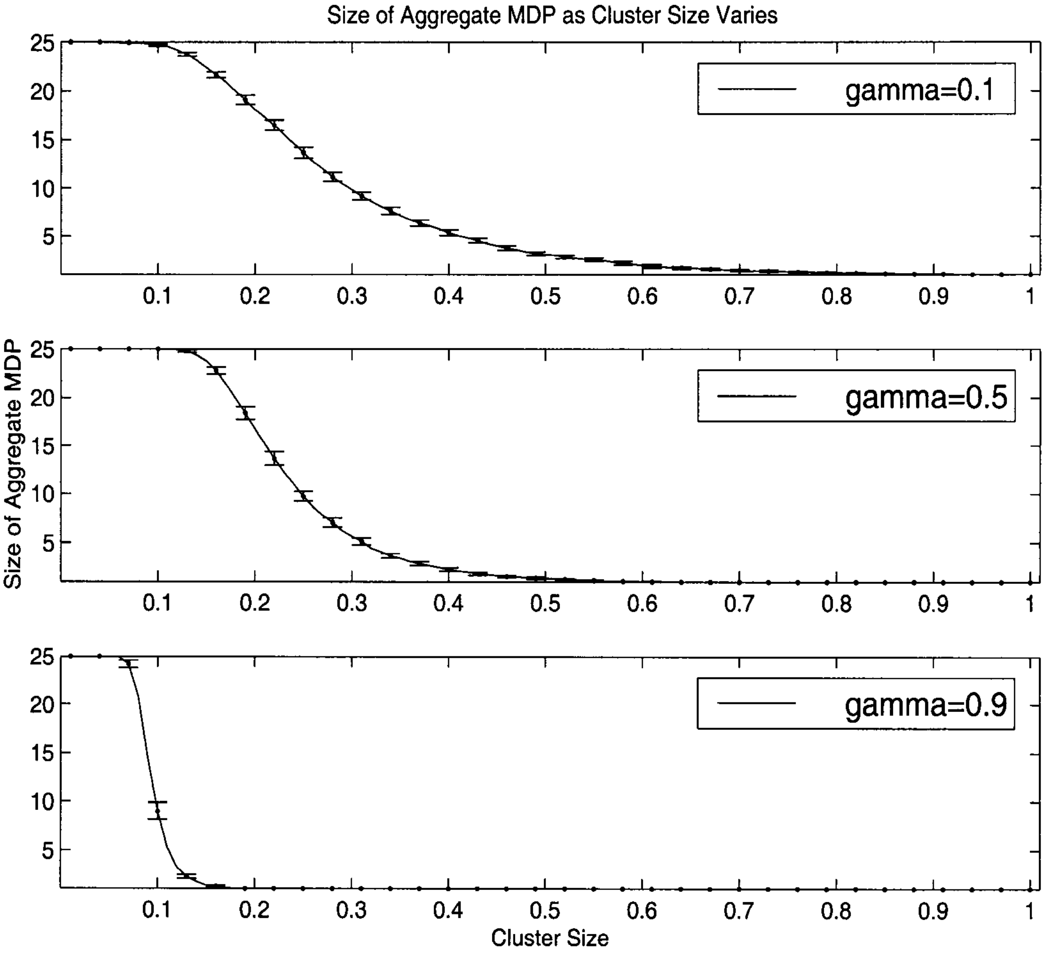

Figure 4.15: Aggregate MDP Sizes for $|S| = 25$, $|A| = 5$, $B = 10$

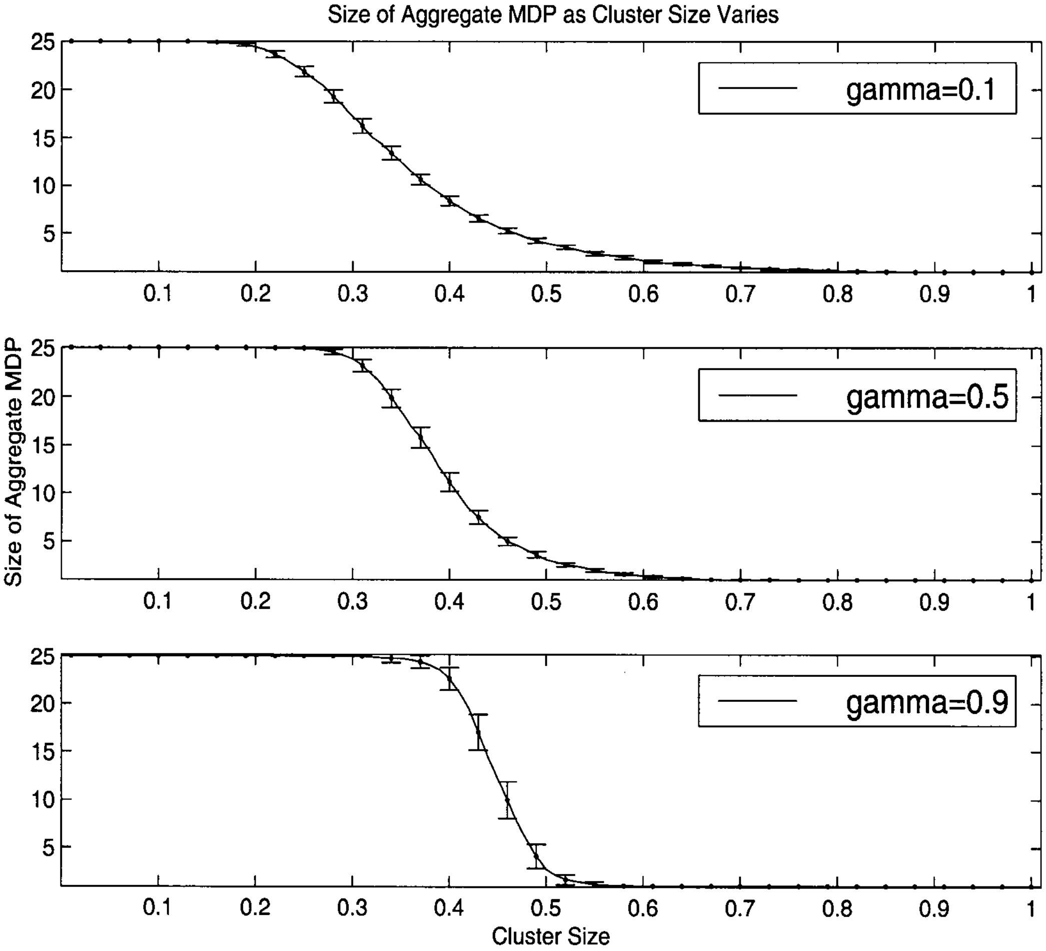

Figure 4.16: Aggregate MDP Sizes for $|S| = 25$, $|A| = 10$, $B = 2$

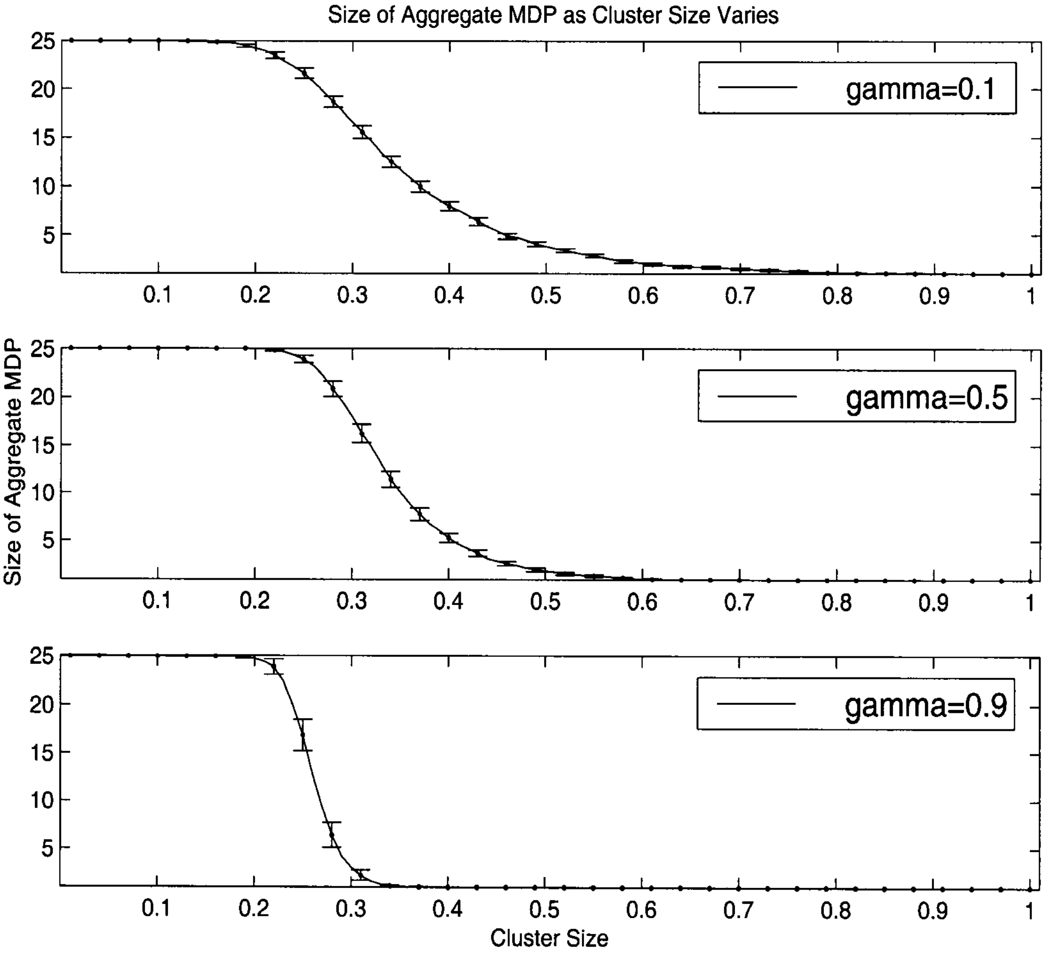

Figure 4.17: Aggregate MDP Sizes for $|S| = 25$, $|A| = 10$, $B = 5$

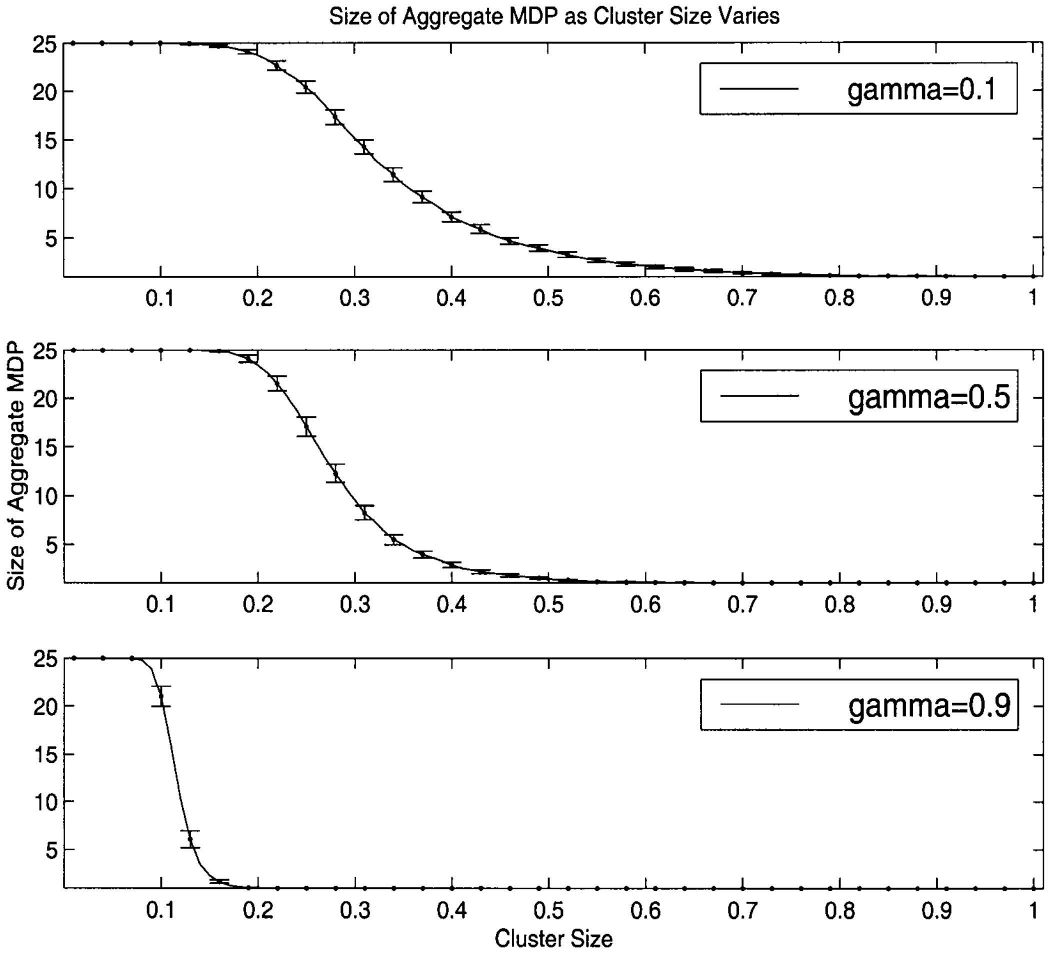

Figure 4.18: Aggregate MDP Sizes for $|S| = 25$, $|A| = 10$, $B = 10$

# Chapter 5

# Conclusion

## 5.1 Summary

In this work we have established the metric analogue of bisimulation for finite MDPs, thereby providing a new tool to reason about such models. As we have seen, such metrics can be used for finite MDP state aggregation. More specifically, the contributions of this thesis are:

- the formulation of bisimulation metrics for discounted reinforcement learning tasks
- a polynomial time algorithm for computing such metrics up to a prescribed degree of accuracy
- bounds relating the values of states to distances assigned by bisimulation metrics
- bounds relating the values of states in a compressed MDP to the values of states in the original MDP, in terms of distances assigned by a

bisimulation metric

## 5.2 Related Work

Much of the recent work on metric based compression of probabilistic systems comes from [4], [23], [24], and [5] and indeed our presentation is based heavily on these.

With respect to compression of MDPs the works closest in nature to our own are those of [8] and [3]. In the former, Givan, Dean, and Greig use an iterative algorithm based on exact bisimulation to aggregate MDPs. Thus, their technique is subject to the same shortcomings, due to bisimulation being too restrictive, discussed in the beginning of chapter 3. In the latter, Givan, Dean, and Leach aggregate MDPs using *$\epsilon$-approximate stochastic bisimulation homogeneous partitions*; these are partitions of the state space in which for each class and each pair of states in the class, the states' respective rewards and their respective transition probabilities to other classes each differ by at most $\epsilon$. This is not aggregation in the usual sense; rather, an $\epsilon$-homogeneous partition is used to define a *bounded Markov decision process* (BMDP) in which reward and transition functions map to closed intervals of real numbers, instead of single values. A modified form of value iteration, known as *interval value iteration*, can then be applied to the BMDP to yield an interval of values bounding the optimal value of a state, for each state in the original MDP. Our method, by contrast, yields bounds on optimal values of states and classes in terms of distances between states in a class, which is tighter than an $\epsilon$-cluster radius.

As a final remark, we note that each of the methods above, unlike our own, have been tailored to factored Markov decision processes (see below).

## 5.3 Future Work

1. **Improved algorithm.** One of the primary advantages of our fixed point presentation of the reinforcement learning bisimulation metrics is that it readily admits a polynomial time algorithm for computing such metrics. Unfortunately, this algorithm is still too slow to be useful in practice. One reason for this is that computing the Kantorovich metric, while strongly polynomial, is computationally intensive. Obviously since our algorithm involves iteratively computing many Kantorovich metrics this contributes to a significant slowdown, especially when compared to simply solving the original MDP directly. In order to improve performance we will either have to find an extremely efficient means of computing the Kantorovich metric, replace it with an easily computable approximation, or find a way of eliminating it altogether from our computations.

2. **Continuous state space MDP.** We briefly mentioned in the introduction that our ultimate goal is to extend bisimulation metrics to handle continuous state space MDPs. In some respects we are half way there, as much of our development of the bisimulation metrics has been done in sufficient generality to hold in a continuous setting.

3. **POMDP.** In many real-world situations a decision maker does not know with absolute certainty the current state of the environment. In-

stead the agent can only make observations about the current state of the environment, observations which typically provide only incomplete information due to noise or uncertainty. One says that the environment is "partially observable". This leads to an extension of the MDP model known as the *partially observable Markov decision process* (POMDP). Formally, a POMDP is simply an MDP with an additional set of observations, $O$, and a probability distribution on $O$ at time $t+1$ given that the system transitioned from unobserved state $s$ at time $t$ to $s'$ at time $t+1$ under action $a$.

Naturally, we would like to extend our bisimulation metrics to handle POMDPs. A well-known result states that for every POMDP $P$ a continuous state space MDP $M_P$ can be constructed such that a solution to one gives a solution to the other. Thus, one way of extending our metrics would be through handling continuous state space models. However, there do exist methods for solving POMDPs directly and it would likewise be desirable to directly extend our metrics to handle POMDPs.

4. **Factored MDP.** Although we have worked exclusively with the finite MDP, the model of choice in practice is a slight variation known as the *factored Markov decision process.* In this model the state space is generated by a finite set of state variables, each taking on finitely many values. Enormous representational savings are made by working with the state variables instead of the explicit state space of state variable vectors, since there are exponentially many of the latter. Moreover, many data structures and algorithms take advantage of the factored

nature of such models to provide even more efficiency.

In order to get the most value out of our bisimulation metrics it is necessary to apply them to factored MDPs. One could immediately do so by applying the metrics to the explicit state space of a factored MDP. This is, however, highly undesirable as one would lose the enormous savings achieved by the factored representation. Instead, we need to redefine such metrics to directly take advantage of the structure underlying factored MDPs.

5. **Action space aggregation.** Our investigation into metric based MDP compression has focused only on reducing the state space. Since large actions spaces are also a limiting factor in RL tasks, a natural extension would be to apply such methods to aggregating action spaces.

# Bibliography

[1] A.G. Barto, R.S. Sutton. *Reinforcement Learning: An Introduction.* MIT Press, 1998. [http://www-anw.cs.umass.edu/~rich/book/the-book.html] (23/10/03)

[2] R. Blute, J. Desharnais, A. Edalat, and P. Panangaden. *Bisimulation for labelled Markov processes.* In Proc. LICS'97, pages 149–159, Warsaw, 1997.

[3] T. Dean, R. Givan, S. Leach. *Model reduction techniques for computing approximately optimal solutions for Markov decision processes.* In Proceedings of UAI-97, pp. 124- 131, 1997. [http://citeseer.nj.nec.com/article/dean97model.html] (10/20/03)

[4] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. *Metrics for labeled Markov systems.* In J.C.M. Baeten and S. Mauw, editors, Proceedings of the 10th International Conference on Concurrency Theory, volume 1664 of Lecture Notes in Computer Science, pages 258–273, Eindhoven, August 1999. SpringerVerlag.

[5] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. *The Metric analogue of weak bisimulation for probabilistic processes.* In the Proceedings of the IEEE Conference on Logic in Computer Science, 2002

[6] R.M. Dudley. *Real Analysis and Probability.* Wadsworth & Brooks/Cole, 1989.

[7] Alison L. Gibbs, Francis Edward Su. 2002. *On Choosing and Bounding Probability Metrics.* [`http://arxiv.org/abs/math.PR/0209021`](28/07/03)

[8] R. Givan, T. Dean and M. Greig. 2001. *Equivalence notions and model minimization in Markov Decision Processes.* Artificial Intelligence, 2003

[9] Jason H. Goto, Mark E. Lewis, and Martin L. Puterman. 2001. *Coffee, Tea, or ...?: A Markov Decision Process Model for Airline Meal Provisioning.* [`http://citeseer.nj.nec.com/goto01coffee.html`] (27/09/03)

[10] M. Hennessy and R. Milner. *Algebraic laws for nondeterminism and concurrency.* J. Assoc. Comput. Mach., 32:137-161, 1985.

[11] Leslie Pack Kaelbling and Terran Lane. 2001. *Approaches to macro decompositions of large Markov decision process planning problems.* [`citeseer.nj.nec.com/lane01approaches.html`] (11/05/03)

[12] L.P. Kaelbling, M.L. Littman, and A.W. Moore. 1996. *Reinforcement learning: a survey.* [`http://www.cs.washington.edu/research/jair/volume4/kaelbling96a-html/rl-survey.html`] (04/09/03)

[13] H. Kautz and R. Sutton. *RandomMDPs.lisp: Random MDPs and POMDPs.* [`http://www.cs.ualberta.ca/~sutton/RandomMDPs.html`] (29/11/03)

[14] D. Kozen. *A Probabilistic PDL.* Journal of Computer and Systems Sciences, 30(2):162-178, 1985

[15] K.G. Larsen, and A. Skou. 1991. *Bisimulation through Probabilistic Testing.* In Information and Computation 94(1):128.

[16] R. Milner. 1989. Communication and Concurrency. Series in Computer Science. Prentice-Hall International.

[17] J.B. Orlin. *A Faster Strongly Polynomial Minimum Cost Flow Algorithm.* Operations Research, vol.41, no.2, pp. 338-50, 1993.

[18] D. Park. 1981. *Concurrency and automata on infinite sequences.* In 5 GI Conference, Karlsruhe, Germany, P. Deussen, ed., vol. 104 of Lecture Notes in Computer Science, Springer-Verlag, 1981, pp. 167–183.

[19] K.R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press, 1967.

[20] Martin L. Puterman. Markov decision processes: discrete stochastic dynamic programming. Wiley series in probability and mathematical statistics. John Wiley & Sons, 1994.

[21] S. T. Rachev, and L. Rüschendorf. *Mass Transportation Problems, Vol. I: Theory.* Springer, Berlin Heidelberg New York, 1998.

[22] D. Ramachandran, L. Rüschendorf. *A general duality theorem for marginal problems.* In Probab. Theory Related Fields 101 (1995), 311-319.

[23] Franck van Breugel and James Worrell. *Towards Quantitative Verification of Probabilistic Transition Systems.* In F. Orejas, P.G. Spirakis and J. van Leeuwen, editors, Proceedings of the 28th International Colloquium on Automata, Languages, and Programming (ICALP), volume 2076 of Lecture Notes in Computer Science, pages 421-432, Crete, July 2001. Springer-Verlag.

[24] Franck van Breugel and James Worrell. *An Algorithm for Quantitative Verification of Probabilistic Transition Systems.* In K.G. Larsen and M. Nielsen, editors, Proceedings of the 12th International Conference on Concurrency Theory (CONCUR), volume 2154 of Lecture Notes in Computer Science, pages 336-350, Aalborg, August 2001. Springer-Verlag.

[25] B. Van Roy. *Neuro-Dynamic Programming: Overview and Recent Trends.* In Handbook of Markov Decision Processes: Methods and Applications, edited by E. Feinberg and A. Shwartz, Kluwer, 2001. [`http://www.stanford.edu/~bvr/psfiles/ndp.ps`] (06/05/03)

[26] C. Villani. 2002 *Topics in Mass Transportation.* [`http://www.math.toronto.edu/hmaroofi/seminar/articles/Vilnotes.ps`](28/07/03)

[27] J. Vygen. *On Dual Minimum Cost Flow Algorithms.* Mathematical Methods of Operations Research 56 (2002), 101-126 Extended abstract

in the Proceedings of the 32nd Annual ACM Symposium on the Theory of Computing (2000), 117-125

[28] Glynn Winskel. The Formal Semantics of Programming Languages: An Introduction. MIT Press, 1993.