Approximate Homomorphisms: A framework for non-exact minimization in Markov Decision Processes

Balaraman Ravindran
Dept. of Computer Science and Engg.
Indian Institute of Technology Madras
Chennai 600 036, TN, India
E-mail: *ravi AT cs.iitm.ernet.in*

Andrew G. Barto
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01002, USA
E-mail: *barto AT cs.umass.edu*

Abstract

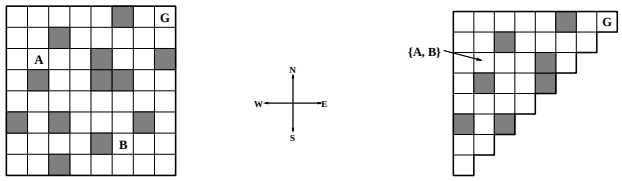
To operate effectively in complex environments learning agents require the ability to selectively ignore irrelevant details and form useful abstractions. In earlier work we explored in detail what constitutes a useful abstraction in a stochastic sequential decision problem modeled as a Markov Decision Process (MDP). We based our approach on the notion of an MDP homomorphism. In this article we look at relaxing the strict conditions imposed earlier and introduce approximate homomorphisms that allow us to construct useful abstract models even when the homomorphism conditions are not met exactly. We also present a result on bounding the loss resulting from this approximation.

1 Introduction

The ability to form abstractions is one of the features that allow humans to operate effectively in complex environments. Researchers in many fields, ranging from various branches of mathematics to social network analysis, also recognize the utility of abstractions and have tried to answer questions such as what is a useful abstraction and how to model abstractions. Abstract representations keep recurring in various guises in the literature. Informally, one can define a good abstraction to be a function of the observable features of a task such that it is a "sufficient statistic".

Determining sufficiency and providing ways of modeling abstractions are well studied problems in AI (e.g., [Amarel, 1968; Knoblock, 1990; Givan *et al.*, 2003]). They are difficult problems when stated in general terms. Therefore, much of the work in this field is specialized to particular classes of problems or specific modeling paradigms. In our work, we focus on Markov decision processes (MDPs), a formalism widely employed in modeling and solving stochastic sequential decision problems.

Our approach to MDP abstraction is based on the notion of *MDP homomorphisms* [Ravindran and Barto, 2002]. This is an extension of machine homomorphisms from the finite state automata (FSA)



(a) (b)

Figure 1: (a) A symmetric gridworld problem. The goal state is G and there are four deterministic actions. State-action pairs (A, E) and (B, N) are equivalent (see text). (b) A reduced model of the gridworld in (a). The state-action pairs (A, E) and (B, N) in the original problem both correspond to the pair $(\{A, B\}, E)$ in the reduced problem.

literature [Hartmanis and Stearns, 1966]. Machine homomorphisms help establish precise correspondences between automata that have similar behavior and identify states that can be aggregated together to derive "smaller" equivalent models. We extend the notion to MDPs by incorporating decision making and stochasticity. But the power of our approach comes from employing *a state-dependent action recoding*. This enables us to apply our results to a wider class of problems and extend existing MDP abstraction frameworks in ways not possible earlier. Our approach to abstraction belongs to the class of algorithms known as model minimization methods and can be viewed as an extension of the MDP minimization framework proposed by Dean and Givan [Givan *et al.*, 2003].

To illustrate the concept of minimization, consider the simple gridworld shown in Figure 1(a). The goal state is labeled G. Taking action E in state A is equivalent to taking action N in state B, in the sense that they go to equivalent states that are both one step closer to the goal. One can say that the state-action pairs (A, E) and (B, N) are equivalent. One can exploit this notion of equivalence to construct a smaller model of the gridworld (Figure 1(b)) that can be used to solve the original problem.

While abstractions that lead to exact equivalences are very useful, they are often difficult to achieve. To apply our approach to real-world problems we need to consider a variety of "relaxed" minimization criteria. For example, in the gridworld in Figure 1 assume that the action E succeeds with probability 0.9 and the action N succeeds with probability 0.8. When actions fail, you stay in the same cell. We could still consider (A, E) and (B, N) equivalent for minimization purposes.

In this article we explore a relaxation of our minimization framework to accommodate *approximate equivalence* of state-action pairs. We use results from [Whitt, 1978] to bound the loss in performance resulting from our approximations. Specifically, we introduce the concept of an *approximate homomorphism* which uses the average behavior of the aggregated states and is particularly useful in learning. In [Ravindran and Barto, 2002] we introduced the concept of a *bounded homomorphism* based on Bounded-parameter MDPs [Givan *et al.*, 2000] and derived loose bounds on the loss of performance resulting from the approximation. Approximate homomorphisms allow us to derive tighter bounds on the loss and also more closely model approximations resulting from online behaviour of a learning or planning agent as opposed to bounded homomorphisms.

In Section 2 we introduce some notation we are using in the work and some background on MDPs. We define MDP homomorphisms and briefly outline the minimization problem in Section 3. Approximate MDP homomorphisms are introduced in Section 4 along with some results on error bounds. We conclude in Section 5 with some discussion and future directions of research.

2 Notation

A *Markov Decision Process* is a tuple $\langle S, A, \Psi, P, R\rangle$, where S is a finite set of states, A is a finite set of actions, $\Psi \subseteq S \times A$ is the set of admissible state-action pairs, $P : \Psi \times S \to [0, 1]$ is the transition probability function with $P(s, a, s')$ being the probability of transition from state s to state s' under action a, and $R : \Psi \to \mathbb{R}$ is the expected reward function, with $R(s, a)$ being the expected reward for performing action a in state s. Let $A_s = \{a|(s, a) \in \Psi\} \subseteq A$ denote the set of actions admissible in state s. We assume that for all $s \in S$, A_s is non-empty.

A *stochastic policy* π is a mapping from Ψ to the real interval $[0, 1]$ with $\sum_{a\in A_s} \pi(s, a) = 1$ for all $s \in S$. For any $(s, a) \in \Psi$, $\pi(s, a)$ gives the probability of picking action a in state s. The *value* of state s under policy π is the expected value of the discounted sum of future rewards starting from state s and following policy π thereafter. The *value function* V^π corresponding to a policy π is the mapping from states to their values under π.

The solution of an MDP is an *optimal policy* $\pi^\star$ that uniformly dominates all other possible policies for that MDP. In other words $V^{\pi^\star}(s) \geq V^\pi(s)$ for all s in S and for all possible π. It can be shown that the value functions for all optimal policies is the same. We denote this *optimal value function* by $V^\star$. It satisfies the *Bellman optimality equation*:

$$V^\star(s) = \max_{a\in A_s} \sum_{s'\in S} P(s, a, s') \left[R(s, a) + \gamma V^\star(s')\right].$$

Given an optimal value function, an optimal policy maybe obtained by behaving *greedily* with respect to it.

Let B be a partition of a set X. For any $x \in X$, $[x]_B$ denotes the block of B to which x belongs. Any function f from a set X to a set Y induces a partition B_f on X, with $[x]_{B_f} = [x']_{B_f}$ if and only if $f(x) = f(x')$. We use the shorthand $[x]_f$ to denote $[x]_{B_f}$. A *partition of an MDP* $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ is a partition of Ψ.

3 MDP Homomorphisms

A homomorphism from a dynamic system $\mathcal{M}$ to a dynamic system $\mathcal{M}'$ is a mapping that preserves $\mathcal{M}$'s dynamics, while in general eliminating some of the details of the full system $\mathcal{M}$. One can think of $\mathcal{M}'$ as a simplified model of $\mathcal{M}$ that is nevertheless a valid model of $\mathcal{M}$ with respect to the aspect's of $\mathcal{M}$'s dynamics that it preserves. The specific definition of homomorphism that we adopt is as follows:

Definition: An *MDP homomorphism* h from an MDP $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ to an MDP $\mathcal{M}' = \langle S', A', \Psi', P', R'\rangle$ is a surjection from Ψ to Ψ', defined by a tuple of surjections $\langle f, \{g_s|s \in S\}\rangle$, with

$h((s,a)) = (f(s), g_s(a))$, where $f : S \to S'$ and $g_s : A_s \to A'_{f(s)}$ for $s \in S$, such that $\forall s, s' \in S, a \in A_s$:

$$P'(f(s), g_s(a), f(s')) = \sum_{s'' \in [s']_f} P(s, a, s''), \quad (1)$$

$$R'(f(s), g_s(a)) = R(s, a). \quad (2)$$

We call $\mathcal{M}'$ the *homomorphic image* of $\mathcal{M}$ under h, and we use the shorthand $h(s,a)$ to denote $h((s,a))$. The surjection f maps states of $\mathcal{M}$ to states of $\mathcal{M}'$, and since it is generally many-to-one, it generally induces nontrivial equivalence classes of states s of M: $[s]_f$. Each surjection g_s recodes the actions admissible in state s of $\mathcal{M}$ to actions admissible in state $f(s)$ of $\mathcal{M}'$. This *state-dependent* recoding of actions is a key innovation of our definition, which we discuss in more detail below. Condition (1) says that the transition probabilities in the simpler MDP $\mathcal{M}'$ are expressible as sums of the transition probabilities of the states of $\mathcal{M}$ that f maps to that same state in $\mathcal{M}'$. This is the stochastic version of the standard condition for homomorphisms of deterministic systems that requires that the homomorphism commutes with the system dynamics [Hartmanis and Stearns, 1966]. Condition (2) says that state-action pairs that have the same image under h have the same expected reward.

The state-dependent action mapping allows us to model symmetric equivalence in MDPs. For example, if $h = \langle f, \{g_s | s \in S\}\rangle$ is a homomorphism from the gridworld of Figure 1(a) to that of Figure 1(b), then $f(A) = f(B)$ is the state marked $\{A, B\}$ in Figure 1(b). Also $g_A(E) = g_B(N) = E$, $g_A(W) = g_B(S) = W$, and so on. A more detailed presentation on modeling symmetries is available in [Ravindran and Barto, 2002].

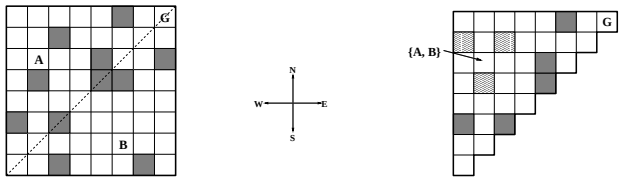
3.1 Minimization

The notion of homomorphic equivalence immediately gives us an MDP minimization framework. In [Ravindran and Barto, 2002] we extended the minimization framework of Dean and Givan [Givan *et al.*, 2003] to include state-dependent action recoding and showed that if two state-action pairs have the same image under a homomorphism, then they have the same optimal value. If $\mathcal{M}'$ is an image of $\mathcal{M}$ under homomorphism $h = \langle f, \{g_s | s \in S\}\rangle$, we also showed that a policy in $\mathcal{M}'$ can *induce* a policy in $\mathcal{M}$ that is closely related. Specifically a policy that is optimal in $\mathcal{M}'$ can be *lifted* to $\mathcal{M}$ to yield an optimal policy in $\mathcal{M}$. Thus we can solve the original MDP by solving a homomorphic image. We define lifting as follows:

Definition: Let π' be a stochastic policy in $\mathcal{M}'$. Then π' *lifted to* $\mathcal{M}$ is the policy $\pi'_{\mathcal{M}}$ such that for any $a \in g_s^{-1}(a')$, $\pi'_{\mathcal{M}}(s,a) = \pi'(f(s), a') \Big/ \left|g_s^{-1}(a')\right|$.

where for any $s \in S$, $g_s^{-1}(a')$ denotes the set of actions that have the same image $a' \in A'_{f(s)}$ under g_s. It is sufficient that $\sum_{a \in g_s^{-1}(a')} \pi'_{\mathcal{M}}(s,a) = \pi'(f(s), a')$, but we use the above definition to make the lifted policy unique.

The goal of minimization is to derive a *minimal image* of the MDP, i.e., a homomorphic image with the least number of admissible state-action pairs. We also adapted an existing minimization algorithm to



(a) (b)

Figure 2: (a) A slightly asymmetric gridworld problem. The goal state is G and there are four deterministic actions. The problem is approximately symmetric about the NE-SW diagonal. (b) A reduced model of the gridworld in (a). The state-action pairs (A, E) and (B, N) in the original problem both correspond to the pair $(\{A, B\}, E)$ in the reduced problem. A solution to this reduced gridworld can be used to derive an approximate solution to the full problem.

find minimal images employing state-action equivalence. Employing state-dependent action recoding allows us to achieve greater reduction in model size than possible with Dean and Givan's framework. For example, the gridworld in Figure 1(a) is minimal if we do not consider state-dependent action mappings.

4 Approximate Equivalence

The MDP homomorphism conditions on which we base our notions of equivalence are very strong conditions and are satisfied only in some restricted classes of problems. Nevertheless in practice we frequently encounter problems for which we can derive useful "approximate" reduced models by employing relaxed notions of equivalence. We construct these approximate models by aggregating together states and actions that differ slightly in their dynamics. For example, consider the gridworld shown in Figure 2(a). This is a slightly modified version of the symmetric grid world from Figure 1(a). While the MDP is more or less symmetric about the NE-SW diagonal as before, there are a few states including A and B that are not symmetric. These differences do not affect the optimal policy for reaching the goal significantly, and we can form a reduced MDP (Figure 2(b)) which is similar to the MDP shown in Figure 1(b). Here we need to treat the lightly shaded states and their neighbours differently, since these are non-symmetric states.

In [Ravindran and Barto, 2002] we introduced the concept of a *bounded homomorphism* to model this approximate minimization. Bounded homomorphisms employed Bounded-parameter MDPs (BMDPs) [Givan *et al.*, 2000] in which the transition probabilities and the rewards are specified as intervals. This allowed us to derive bounds on the optimal value function of the original MDP. Since BMDPs actually specify a family of MDPs, most of which do not correspond to any of the scenarios being approximated, the bounds derived are very loose. In this section we introduce the concept an *approximate homomorphism* that uses the average behavior of the aggregated states and is particularly useful in learning. We use earlier work by [Whitt, 1978] to derive bounds on the loss of performance resulting from the approximation.

4.1 Approximate Homomorphisms

In many circumstances we can aggregate together states and actions that have slightly different dynamics to form a reduced model. The most straight forward choice for the dynamics of the reduced model is a "weighted" average of the dynamics of the state-action pairs that belong to the same equivalence class. Formally we define an approximate homomorphism as follows:

Definition: An *approximate MDP homomorphism* h from an MDP $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ to an MDP $\mathcal{M}' = \langle S', A', \Psi', P', R'\rangle$ is a surjection from Ψ to Ψ', defined by a tuple of surjections $\langle f, \{g_s | s \in S\}\rangle$, with $h((s,a)) = (f(s), g_s(a))$, where $f : S \to S'$ and $g_s : A_s \to A'_{f(s)}$ for $s \in S$, such that for all s, s' in S and $a \in A_s$:

$$P'(f(s), g_s(a), f(s')) = \sum_{(q,b)\in[(s,a)]_h} w_{qb} \sum_{s''\in[s']_f} P(q, b, s'') \tag{3}$$

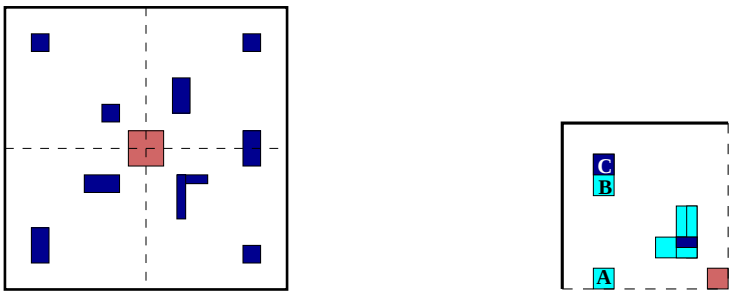
$$R'(f(s), g_s(a)) = \sum_{(q,b)\in[(s,a)]_h} w_{qb}\, R(q, b). \tag{4}$$

where, w_{qb} is a weighting factor for the pair (q, b), with $\sum_{(q,b)\in[(s,a)]_h} w_{qb} = 1$. We call $\mathcal{M}'$ the *approximate homomorphic image* of $\mathcal{M}$ under h. To determine the transition probability $P'(f(s), g_s(a), f(s'))$ in $\mathcal{M}'$ we first compute the block transition probability from each element of $[(s,a)]_h$ to the block $[s']_f$. Then we set the transition probability to be the weighted average of these block transition probabilities. Note that if h is a homomorphism, then the induced partition satisfies the stochastic version of the SP property [Hartmanis and Stearns, 1966] and each of the block transition probabilities we compute above are equal to one another. We do a similar computation for the reward function as well.

When we employ such approximate reduced models to do planning or learning, the appropriate aggregate dynamics to employ is a weighted average of the dynamics of the state-action pairs that belong to a given equivalence class, the weights, w_{qb}, being determined by the frequency with which each member of the class is encountered in the course of the solution process. Usually, while learning with online experience it is sufficient to specify only the state, action and reward spaces of the image MDP and the trajectories through state-action space the agent experiences implicitly induce the transition probabilities. In such cases, the induced transition probabilities of the image MDP will account for the frequency of visitation. In the absence of additional knowledge about the problem, a useful heuristic is to consider a simple average of the aggregated dynamics. In this case we set $w_{qb} = \frac{1}{|[(s,a)]_h|}$.

Example of an Approximate Homomorphism

Consider the MDP shown in Figure 3(a). This represents a spatial navigation task. The goal is to reach the set of shaded states in the center of the environment. The darker regions are obstacles, while the clear regions are open space. Consider the four quadrants formed by the dotted lines in the figure, it is clear that the environment is more or less symmetric. An approximate homomorphic image of this MDP can be formed as shown in Figure 3(b). Once again, the clear regions are open space and the dark regions like C are obstacles. The lightly shaded regions like A and B use aggregate dynamics



(a) (b)

Figure 3: (a) A spatial navigation problem. The goal is to reach the shaded region in the center. The environment is approximately symmetric about the dotted lines. (b) An approximate homomorphic image of the task in (a). Transitions into the lightly shaded regions are determined by aggregate dynamics. See text for more details.

as described above. In particular, if we assume that the original problem was deterministic, then the probability of being able to move into region A is one half and that of being able to move into region B is three quarters.

4.2 Bounding the Approximation Loss

With the relaxation of the homomorphism conditions we lose some of the guarantees we established earlier. In particular the *optimal value equivalence theorem* is no longer guaranteed to hold. A policy that is optimal in an approximate homomorphic image is not necessarily optimal when lifted to the original MDP. But if the approximation is a reasonable one, the lifted policy is not too far from the optimal. We would like to bound the "distance" between the true optimal policy and the policy lifted from the image. We do this by deriving an upper limit on the maximum difference between the optimal value function in the original MDP and the value function of the lifted policy.

We adopt results from [Whitt, 1978] to derive this bound. Whitt explored the issue of approximation and abstraction in the *contraction mapping* formulation of a dynamic program due to [Denardo, 1967]. This formulation is a generalization of MDPs, stochastic games and other sequential decision making paradigms. Whitt explores the issue of approximating a dynamic program from the point of optimal value preservation and considers state-action equivalence, along with state-action value functions. He derives precise conditions for when an image accurately captures the optimal values of the original dynamic program and also looks at sequence of approximations that in the limit converge to an exact image. He also derives bound on the loss in the optimal value function when the image is an approximation. He specializes some of the results to stochastic sequential decision problems, from which we can derive the equivalent results for MDPs. Here we further specialize his work to our formulation of an approximate homomorphism.

The bounds depend on the differences in the resulting aggregate parameters and the actual parameters. Let $h = \langle f, \{g_s | s \in S\}\rangle$ be an approximate homomorphism from an MDP $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ to an

MDP $\mathcal{M}' = \langle S', A', \Psi', P', R' \rangle$. We define the following quantities:

$$\begin{aligned} K_r &= \max_{\substack{s \in S \\ a \in A_s}} \mid R(s,a) - R'(f(s), g_s(a)) \mid \\ K_p &= \max_{\substack{s \in S \\ a \in A_s}} \sum_{[s_1]_f \in B_f} \left| \sum_{t \in [s_1]_f} P(s,a,t) - P'(f(s), g_s(a), f(s_1)) \right| \\ \delta_{r'} &= \max_{\substack{s' \in S' \\ a' \in A_{s'}}} R'(s',a') - \min_{\substack{s' \in S' \\ a' \in A_{s'}}} R'(s',a') \end{aligned}$$

where K_r is the maximum difference between the aggregate block reward and the actual reward, K_p is the maximum difference between the actual block transition probabilities and the aggregate transition probabilities and δ'_r is the range of the reward function in the image MDP. Then the following theorem holds:

Theorem: Let $\pi'^\star$ be an optimal policy in $\mathcal{M}'$ and $\pi'^\star_M$ be that policy lifted to $\mathcal{M}$. Let γ be the discount factor. Then:

$$\left\| V^\star - V^{\pi'^\star_M} \right\| \leq \frac{2}{1-\gamma} \left(K_r + \frac{\gamma}{1-\gamma} \delta_{r'} \frac{K_p}{2} \right)$$

.

Proof: Let $h = \langle f, \{g_s | s \in S\} \rangle$ be the approximate homomorphism from $\mathcal{M}$ to the image $\mathcal{M}'$. Let $V^{\pi'^\star}$ be the optimal value function in $\mathcal{M}'$. Let $\tilde{V}^{\pi'^\star}$ be the function constructed by *lifting* $V^{\pi'^\star}$ to $\mathcal{M}$, i.e., $\tilde{V}^{\pi'^\star}(s) = V^{\pi'^\star}(f(s))$. Note that $\tilde{V}^{\pi'^\star}$ is not necessarily the same function as $V^{\pi'^\star_M}$ since h is only an approximate homomorphism. Let

$$Q(s,a,V) = R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V(s'),$$

for some real valued function V on S. From Theorem 6.1 of [Whitt, 1978] we have the following:

$$K(V^{\pi'^\star}) = \max_{\substack{a \in A_s \\ s \in S}} \left| Q(s,a,\tilde{V}^{\pi'^\star}) - Q(f(s), g_s(a), V^{\pi'^\star}) \right| \leq K_r + \frac{\gamma}{1-\gamma} \delta_{r'} \frac{K_p}{2}.$$

Since $\mathcal{M}$ and $\mathcal{M}'$ are MDPs, we employ corollary (b) of Theorem 6.1 here. From the corollary to Lemma 3.1 of [Whitt, 1978] we have:

$$\left\| V^\star - V^{\pi'^\star_M} \right\| \leq \frac{2}{1-\gamma} K(V^{\pi'^\star}).$$

Since $V^{\pi'^\star}$ is optimal in $\mathcal{M}'$, we omit the terms that arise due to deviation from optimality of the value function. We get the desired result by combining the above two results. □

Here the distance between the value functions is measured using the max-norm, i.e., $\|V_1 - V_2\| = \max_{s \in S} |V_1(s) - V_2(s)|$. Thus our Theorem allows us to bound the maximum deviation from the true optimal function that results from using a particular approximate homomorphic image. This result holds only for values of $\gamma < 1$. When γ is 1, it is possible to construct examples where the error is

unbounded. For small value of γ the overall error depends more on the difference in the immediate reward, since the second term within the parenthesis is small. This is not surprising, since small γ leads to more myopic optimal policies. Similarly for large values of γ the error depends more on the differences in the transition probabilities and the range of the rewards function, since these quantities affect the long term return.

While the derivation of the bound does not depend on the details of the averaging method used, the bounds themselves could vary, since how we average influences the values of K_p, K_r and $\delta_{r'}$. Therefore these bounds can be computed beforehand if we are using simple averages or some fixed weighted averaging scheme. If we want to use the visitation frequencies in order to derive our reduced MDPs, we need to dynamically recompute our bounds as we gather more information.

5 Discussion and Future Work

The definition of an approximate homomorphism we introduced here is very inclusive. In fact, it is possible to define an approximate homomorphism from any MDP to any other arbitrary MDP. But the bounds given by our Theorem will be very loose and there is no practical utility in defining such homomorphisms. An useful approximate homomorphism should be one in which the values of K_r and K_p are "reasonable". One measure of usefulness is to check if the loss in performance that results from lifting the solution of the approximate image to the original problem is acceptable.

Our Theorem gives an upper bound on the loss of performance due to the approximation. The lower bound even when we use an approximate homomorphic image is zero. In other words, it is still possible to recover the optimal solution of the original problem by lifting the solution of an approximate homomorphic image. Such a situation arises due to the fact that optimal policies when defined as acting greedily with respect to the optimal value functions are sensitive only to the relative ordering of the values. In fact, we need to identify correctly only the action with the highest value in each state. This is possible in many situations when we apply approximate homomorphic images and enhances the utility of such images. Such results were observed in several experiments employing approximate homomorphisms reported elsewhere [Ravindran and Barto, 2002; 2003]. One future direction of research is to arrive at bounds that are better indicators of optimal policy performance.

[Givan *et al.*, 2000] developed the notion of a BMDP while studying approximate minimization in Dean and Givan's minimization framework. They base their work on related formulations of *MDPs with imprecise parameters* (e.g., [Satia and Lave, 1973]) from operations research. Givan et al. also investigate in detail the question of constructing reduced BMDP models of given MDPs. They conclude that it is not usually possible to specify a unique reduced BMDP model, and that we have to resort to some heuristic to choose between several equally viable alternatives. They also show that we cannot guarantee that there is a heuristic which in all cases will lead to the best possible BMDP, i.e., one that gives the best bounds on the value functions and the smallest reduced models. The utility of the bounded approximate homomorphism formulation is as a tool for deriving a priori bounds, loose bounds in many cases, on the loss in performance when we employ a particular abstraction.

[Whitt, 1978] uses a notion of approximation similar to our definition of approximate homomorphism. His motivation was to develop an abstraction framework for dynamic programs. He outlines a method

to derive homomorphic images of dynamic programs by successively refining the approximate image so that the error bounds become tighter. [Kim and Dean, 2001] have developed a similar method for MDPs. They also successively construct better approximate images of a given MDP, but the criterion they use to refine the image is the actual performance of the image MDPs' solutions when lifted to the original MDP. We believe that this is a more promising direction for developing iterative algorithms to finding minimal images of an MDP. Currently we are working on combining partial knowledge of symmetries of the system with the above framework to design more efficient minimization algorithms.

Acknowledgments

We wish to thank Eric Denardo for pointing us to Ward Whitt's work on approximate dynamic programs. This material is partly based upon work supported by the National Science Foundation under Grant No. ECS-0218125 to Andrew G. Barto and Sridhar Mahadevan. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[Amarel, 1968] S Amarel. On representations of problems of reasoning about actions. *Machine Intelligence*, 3:131–137, 1968.

[Denardo, 1967] E. V. Denardo. Contraction mappings in the theory underlying dynamic programming. *SIAM Review*, 9:165–177, 1967.

[Givan *et al.*, 2000] R. Givan, S. Leach, and T. Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122:71–109, 2000.

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

[Hartmanis and Stearns, 1966] J. Hartmanis and R. E. Stearns. *Algebraic Structure Theory of Sequential Machines*. Prentice-Hall, Englewood Cliffs, NJ, 1966.

[Kim and Dean, 2001] K-E. Kim and T. Dean. Solving factored MDPs via non-homogeneous partitioning. In *Proceedings of the 17th IJCAI*, pages 747–752, 2001.

[Knoblock, 1990] C. A. Knoblock. Learning abstraction hierarchies for problem solving. In *Proceedings of the 8th AAAI*, volume 2, pages 923–928, Boston, MA, 1990. MIT Press.

[Ravindran and Barto, 2002] B. Ravindran and A. G. Barto. Model minimization in hierarchical reinforcement learning. In S. Koenig and R. C. Holte, eds, *Proceedings of SARA 2002, LNAI 2371*, pages 196–211, New York, NY, August 2002. Springer-Verlag.

[Ravindran and Barto, 2003] B. Ravindran and A. G. Barto. Relativized options: Choosing the right transformation. In *Proceedings of the 20th ICML*, pages 608–615, Menlo Park, CA, August 2003.

[Satia and Lave, 1973] J. K. Satia and R. E. Lave. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21:728–740, 1973.

[Whitt, 1978] W. Whitt. Approximations of dynamic programs I. *Mathematics of Operations Research*, 3(3):231–243, 1978.