Model Minimization in Hierarchical Reinforcement Learning

Balaraman Ravindran and Andrew G. Barto

Department of Computer Science,
University of Massachusetts,
Amherst, MA 01003, USA
{ravi,barto}@cs.umass.edu

Abstract. When applied to real world problems Markov Decision Processes (MDPs) often exhibit considerable implicit redundancy, especially when there are symmetries in the problem. In this article we present an MDP minimization framework based on homomorphisms. The framework exploits redundancy and symmetry to derive smaller equivalent models of the problem. We then apply our minimization ideas to the options framework to derive relativized options—options defined without an absolute frame of reference. We demonstrate their utility empirically even in cases where the minimization criteria are not met exactly.

1 Introduction

Researchers in artificial intelligence (AI) and in particular machine learning (ML) have long recognized that extending AI and ML approaches to more complex real-world domains requires incorporating the ability to handle and form various abstractions, both temporal and spatial. In this article we present a Markov decision processes (MDP) minimization framework we developed earlier [15] that allows us to abstract away redundancy in the problem definition. We then apply these ideas to hierarchical Reinforcement Learning (RL). Our framework is an extension of a MDP minimization framework developed by Dean and Givan [4, 6].

Model minimization methods attempt to abstract away redundancy in an MDP model and derive an "equivalent" smaller model. To illustrate model minimization, consider the simple gridworld shown in Figure 1(a). The goal state is labelled G. The gridworld is symmetric about the NE-SW diagonal. Hence 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 one step closer to the goal. One can say that the state-action pairs (A, E) and (B, N) are equivalent. We can exploit this notion of equivalence to construct a smaller model of the gridworld, one that can be used to derive a solution to the original problem. Such a reduced gridworld is shown in Figure 1(b).

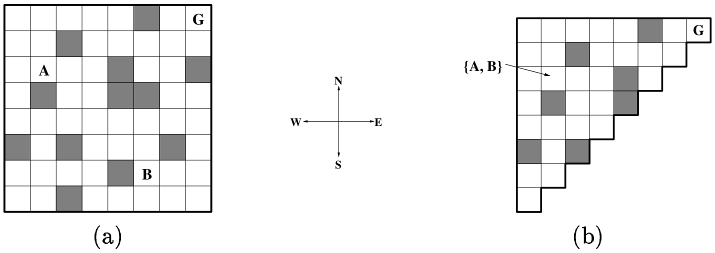


Fig. 1. (a) A symmetric gridworld problem. The goal state is G and there are four deterministic actions. States A and B are equivalent in the sense described in the text. (b) A reduced model of the gridworld in (a). The states A and B in the original problem correspond to the single state $\{A, B\}$ in the reduced problem. A solution to this reduced gridworld can be used to derive a solution to the full problem.

We base our approach to MDP minimization on the notion of *MDP homomorphisms.* This is an extension of machine homomorphisms from finite state automata (FSA) literature [9]. We extend the notion to MDPs by incorporating decision making and stochasticity. The key novelty in our approach is the extension of notions of equivalence to state-action pairs. This enables us to apply our results to a wider class of problems and extend existing MDP minimization frameworks in ways not possible earlier. Specifically, by employing group theoretic concepts we show that our extended minimization framework can abstract away symmetries in an MDP model.

The minimization framework we develop for MDPs can be employed readily by RL algorithms for spatial abstraction. The options framework [17] enables RL algorithms to employ temporal abstractions in the form of temporally extended actions, or *options.* Extending our algebraic framework to a hierarchical RL setting such as the options framework opens up additional possibilities. In this paper we introduce *relativized options*, an extension to the options framework based on "partial" MDP homomorphisms that allows us to define option policies without an absolute frame of reference and hence widens the applicability of an option, enables greater knowledge transfer across tasks and more efficient use of experience. We also investigate employing relativized options in cases where the abstraction conditions are not satisfied exactly. We introduce approximate homomorphisms that model such scenarios using the notion of bounded-parameter MDPs [7].

In the next section we present some notation we will be using. In Section 3 we outline our model minimization framework and state some results. We also show how our framework can exploit symmetries of MDPs. In Section 4 we introduce relativized options and present some experimental results. In Section 5 we define approximate homomorphisms and empirically demonstrate their usefulness. We conclude with a discussion on related work and some 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. We assume that the rewards are bounded. 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 solution of an MDP is an *optimal policy* $\pi^\star$ that uniformly dominates all other possible policies for that MDP.

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 $f(x) = f(x')$. Let B be a partition of $Z \subseteq X \times Y$, where X and Y are arbitrary sets. The *projection of* B *onto* X is the partition $B|X$ of X such that for any $x, x' \in X$, $[x]_{B|X} = [x']_{B|X}$ if and only if every block of B containing a pair in which x (x') is a component also contains a pair in which x' (x) is a component. A *partition of an MDP* $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ is a partition of Ψ. Given a partition B of $\mathcal{M}$, the *block transition probability of* $\mathcal{M}$ is the function $T : \Psi \times B|S \to [0,1]$ defined by $T(s,a,[s']_{B|S}) = \sum_{s'' \in [s']_{B|S}} P(s,a,s'')$. In other words, $T(s,a,[s']_{B|S})$ is the probability of transiting from state s to some state in the block $[s']_{B|S}$ (i.e. the block to which state s' belongs) under action a.

An option (or a temporally extended action) [17] in an MDP $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ is defined by the tuple $O = \langle \mathcal{I}, \pi, \beta\rangle$, where the initiation set $\mathcal{I} \subseteq S$ is the set of states in which the option can be invoked, π is the option policy, and the termination function $\beta : S \to [0,1]$ gives the probability of the option terminating in any given state. The option policy can in general can be a mapping from arbitrary sequences of state-action pairs (or histories) to action probabilities.

3 MDP Homomorphisms

In this article we present a formalism that captures the intuitive notion of equivalence illustrated in Figure 1. In Figure 1(a), we consider states A and B equivalent, since for every action in A that puts you in a state a certain distance from the goal, there is an action in B that takes you to an equivalent state at the same distance from the goal. More generally, in an MDP $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ we consider state s_1 equivalent to state s_2 if for every action available in s_1, there is some action in s_2 that results in similar behavior with respect to the transition structure P and vice versa. We also require that the actions be equivalent with respect to the reward function R. We can then derive a simpler model $\mathcal{M}'$ of $\mathcal{M}$ by aggregating together *blocks* of equivalent states. In other words, $\mathcal{M}'$ is a simpler model of $\mathcal{M}$ if there exists a transformation from $\mathcal{M}$ to $\mathcal{M}'$ that preserves the transition and reward structure and maps equivalent states in $\mathcal{M}$ to the same state in $\mathcal{M}'$, and equivalent actions in $\mathcal{M}$ to the same action in $\mathcal{M}'$. An MDP homomorphism from $\mathcal{M}$ to $\mathcal{M}'$ is such a transformation. Formally, we define it as:

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:

$$P'(f(s), g_s(a), f(s')) = T(s, a, [s']_{B_h|S}),\ \forall s, s' \in S, a \in A_s \tag{1}$$

$$R'(f(s), g_s(a)) = R(s, a),\ \forall s \in S, a \in A_s \tag{2}$$

We call $\mathcal{M}'$ the *homomorphic image* of $\mathcal{M}$ under h. We use the shorthand $h(s,a)$ to denote $h((s,a))$. From condition (1) we can see that state-action pairs that have the same image under h have the same block transition behavior in $\mathcal{M}$, i.e., the same probability of transiting to any given block of states with the same image under f. Condition (2) says that state-action pairs that have the same image under h have the same expected reward. This definition of a MDP homomorphism leads to the following definition of equivalence of states and state-action pairs.

Definition: State action pairs (s_1, a_1) and $(s_2, a_2) \in \Psi$ are *equivalent* if there exists a homomorphism h of $\mathcal{M}$ such that $h(s_1, a_1) = h(s_2, a_2)$. States s_1 and $s_2 \in S$ are *equivalent* if i) for every action $a_1 \in A_{s_1}$, there is an action $a_2 \in A_{s_2}$ such that (s_1, a_1) and (s_2, a_2) are equivalent, and ii) for every action $a_2 \in A_{s_2}$, there is an action $a_1 \in A_{s_1}$, such that (s_1, a_1) and (s_2, a_2) are equivalent.

Thus the surjection f maps equivalent states of $\mathcal{M}$ onto the same image state in $\mathcal{M}'$, while g_s is a *state dependent* mapping of the actions in $\mathcal{M}$ onto image actions in $\mathcal{M}'$. 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 policy in $\mathcal{M}'$ *induces* a policy in $\mathcal{M}$ and the following describes how to derive such an induced policy.

Definition: Let $\mathcal{M}'$ be an image of $\mathcal{M}$ under homomorphism $h = \langle f, \{g_s | s \in S\}\rangle$. 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. 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 $a \in g_s^{-1}(a')$, $\pi_{\mathcal{M}}(s, a) = \pi(f(s), a') \Big/ \left|g_s^{-1}(a')\right|$.

Note: 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.

Theorem 1: Let $\mathcal{M}' = \langle S', A', \Psi', P', R' \rangle$ be the image of $\mathcal{M} = \langle S, A, \Psi, P, R \rangle$ under the homomorphism $h = \langle f, \{g_s | s \in S\} \rangle$. If $\pi^\star$ is an optimal policy for $\mathcal{M}'$, then $\pi^\star_{\mathcal{M}}$ is an optimal policy for $\mathcal{M}$.[1]

Theorem 1 establishes that an MDP can be solved by solving one of its homomorphic images. To achieve the most impact, we need to derive a smallest homomorphic image of the MDP, i.e., an image with the least number of admissible state-action pairs. The following definition formalizes this notion.

Definition: An MDP $\mathcal{M}$ is a *minimal MDP* if for every homomorphic image $\mathcal{M}'$ of $\mathcal{M}$, there exists a homomorphism from $\mathcal{M}'$ to $\mathcal{M}$. A *minimal image* of an MDP $\mathcal{M}$ is a homomorphic image of $\mathcal{M}$ that is also a minimal MDP.

The model minimization problem can now be stated as: "find a minimal image of the given MDP". Since this can be computationally prohibitive, we frequently settle for a reasonably reduced model, even if it is not a minimal MDP. This minimization framework extends the approach proposed by Dean and Givan [4, 6]. They employ *stochastic bisimulations* [12] on state sets of MDPs and do not consider state-action equivalence. If we restrict homomorphisms to only the state set, our approach is equivalent to theirs in terms of the reductions achieved. The theoretical results established in their framework hold, with suitable modifications, in our framework also. Specifically, by incorporating our extended definitions of equivalence, we can extend their algorithm for computing minimal models of MDPs to compute minimal models as defined above. Employing state-action equivalence 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 irreducible if we consider state equivalence alone. We also explicitly model symmetries of MDPs in our framework.

3.1 Symmetries of MDPs

We formalize the notion of MDP symmetries employing group theoretic concepts and show that abstracting symmetries is a special case of the minimization procedure we developed above.

Definition: An MDP homomorphism $h = \langle f, \{g_s | s \in S\} \rangle$ from MDP $\mathcal{M} = \langle S, A, \Psi, P, R \rangle$ to MDP $\mathcal{M}' = \langle S', A', \Psi', P', R' \rangle$ is an *MDP isomorphism* from $\mathcal{M}$ to $\mathcal{M}'$ if and only if f and g_s, $s \in S$, are bijective. $\mathcal{M}$ is said to be *isomorphic* to $\mathcal{M}'$ and vice versa. An MDP isomorphism from an MDP $\mathcal{M}$ to itself is an *automorphism* of $\mathcal{M}$.

Intuitively one can see that automorphisms can be used to describe symmetries in a problem specification. In the gridworld example of Figure 1, a reflection of

[1] The proofs of the various theorems are presented in ref. [15].

the states about the NE-SW diagonal and a swapping of actions N and E and of actions S and W is an automorphism. It is easy to see that this mapping captures the equivalence discussed earlier.

Definition: The set of all automorphisms of an MDP $\mathcal{M}$, denoted by Aut$\mathcal{M}$, forms a group under composition of homomorphisms. This group is the *symmetry group* of $\mathcal{M}$.

Let $\mathcal{G}$ be a subgroup of Aut$\mathcal{M}$ denoted by $\mathcal{G} \leq \text{Aut}\mathcal{M}$. The subgroup $\mathcal{G}$ induces a partition $B_{\mathcal{G}}$ of Ψ: $[(s_1, a_1)]_{B_{\mathcal{G}}} = [(s_2, a_2)]_{B_{\mathcal{G}}}$ if and only if there exists $h \in \mathcal{G}$ such that $h(s_1, a_1) = (s_2, a_2)$. Since $\mathcal{G}$ is a subgroup, this implies that there exists $h^{-1} \in \mathcal{G}$ such that $h^{-1}(s_2, a_2) = (s_1, a_1)$.

Theorem 2: Let $\mathcal{G} \leq \text{Aut}\mathcal{M}$ be a group of automorphisms of $\mathcal{M} = \langle S, A, \Psi, P, R \rangle$. There exists a homomorphism $h^{\mathcal{G}}$ from $\mathcal{M}$ to some $\mathcal{M}'$, such that the partition induced by $h^{\mathcal{G}}$, $B_{h^{\mathcal{G}}}$, is the same partition as $B_{\mathcal{G}}$.

The image of $\mathcal{M}$ under $h^{\mathcal{G}}$ is called the *$\mathcal{G}$-reduced image* of $\mathcal{M}$ and if $\pi^\star$ is an optimal policy for some $\mathcal{G}$-reduced image of MDP $\mathcal{M}$, then $\pi^\star_{\mathcal{M}}$ is an optimal policy for $\mathcal{M}$. Frequently the Aut$\mathcal{M}$-reduced model of an MDP is a minimal image. We can take advantage of structure inherent in a symmetry group and the induced partition in developing efficient minimization algorithms.

Ours is not the first work to study symmetries of MDPs. Zinkevich and Balch [19] define symmetries employing equivalence relations on the state-action pairs of an MDP. They do not make connections to group theoretic concepts or to minimization algorithms. They show that the optimal action-value function of a symmetric system is symmetric and suggest that the action-value function entries be duplicated. They also study in some detail symmetries that arise in multi-agent systems.

4 Relativized Options

It is often the case that both conditions of a homomorphism do not hold for the entire Ψ space of an MDP but only over parts of it. For example, consider the problem of navigating in the gridworld environment shown in Figure 2(a). The goal is to be in the central corridor after collecting all the objects in the world. A more complete description of the task is provided in Section 4.1. The entire gridworld as such is irreducible. But each of the rooms in the world are equivalent to one another and simple transformation such as reflections and rotations map them onto each other. Thus we can create a “partial” homomorphic image of this environment, shown in Figure 2(b), with the homomorphic conditions holding only for the states in the rooms and not in the corridor. The states in which the homomorphic conditions do not hold get mapped to a “catch all” absorbing state, shown as a dark oval. All actions from these states get mapped to an absorbing action in the image MDP. Formally we can define a partial homomorphism as follows:

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

$$P'(f(s), g_s(a), f(s')) = T(s, a, [s']_{B_h|S}),\ \forall s \in f^{-1}(S'), s' \in S, a \in A_s \tag{3}$$

$$P'(\tau, \alpha, \tau) = 1.0 \tag{4}$$

$$R'(f(s), g_s(a)) = R(s, a),\ \forall s \in f^{-1}(S'), a \in A_s \tag{5}$$

We call $\mathcal{M}'$ the *partial* homomorphic image of $\mathcal{M}$ under h. The state τ is an absorbing state in $\mathcal{M}'$ with one action α that transitions to τ with probability 1. The homomorphism conditions hold only for states that do not map to τ. All the actions in states that map to τ, map to α. Lifting policies defined in $\mathcal{M}'$ yield policy fragments in $\mathcal{M}$, with action probabilities specified only for elements in the support of h, i.e., $h^{-1}(\Psi')$. In Figure 2, τ corresponds to the state represented as a black oval and α is indicated by the solid arrow. All state-action pairs, with the state component in the corridor, map to (τ, α) under the partial homomorphism. We can extend MDP minimization algorithms to find partial homomorphic images by suitably restricting the search for homomorphisms to a subset of Ψ. One approach to taking advantage of partial homomorphisms is to combine our minimization framework with hierarchical learning approaches.

The *options framework* is a hierarchical learning framework introduced by Sutton, Precup and Singh [17]. Options are temporally extended actions that take multiple time steps to complete. A class of options, known as "sub-goal" options, are defined as policy fragments to achieve a certain sub-goal or accomplish a certain sub-task [14]. Frequently, sub-goal options satisfy the Markov property and the option policy is defined as a map from some subset of Ψ to action probabilities. In such instances it is possible to specify the the desired sub-goal of the option and to implicitly define the option policy as the solution to an *option MDP*.

The option MDP corresponding to a sub-goal option O is given by $\mathcal{M}_O = \langle S' \cup \{\tau\}, A' \cup \{\alpha\}, \Psi', P', R_O\rangle$, where $S' \subseteq S$, is the states in which the option policy needs to be defined, τ, is an absorbing state representing the states in $S - S'$, $A' = A$, $\Psi' = \{(s, a) | (s, a) \in \Psi, s \in S'\} \cup \{(\tau, \alpha)\}$, $P'(s, a, s') = P(s, a, s')$, if $(s, a) \in \Psi'$, $s' \in S'$, $P'(\tau, \alpha, \tau) = 1$, and $P'(s, a, \tau) = \sum_{s' \notin S'} P(s, a, s')$ for all (s, a) in Ψ' and R_O is a reward function chosen depending on the sub-task O represents. We refer to the states in which the option policy is defined, S' in this case, as the *domain* of the option. We can also learn the option policy online by learning a solution to the option MDP. Such an approach is particularly useful when sub-goals are easy to identify but developing policies to achieve such sub-goals are non-trivial.

In the gridworld in Figure 2(a), an option that accomplishes the task of collecting an object and leaving room 1 can be defined as a solution to the MDP in Figure 2(b), with the appropriate reward function. We can define similar options for each of the rooms in the world. Formally we define a Markov sub-goal option as follows:

Definition: A Markov sub-goal option of an MDP $\mathcal{M}$ is defined by $O = \langle \mathcal{M}_O, \mathcal{I}, \beta \rangle$, where $\mathcal{I} \subseteq S$ is the initiation set of the option, $\beta : S \rightarrow [0, 1]$, is the termination function and $\mathcal{M}_O = \langle S' \cup \{\tau\}, A' \cup \{\alpha\}, \Psi', P', R_O \rangle$ is the option MDP.

The option policy π is obtained by solving $\mathcal{M}_O$, treating it as an episodic task [16] with the possible initial states of the episodes given by $\mathcal{I}$ and the termination of each episode determined by the option's termination function β.

As is evident, the option MDP $\mathcal{M}_O$ is a partial homomorphic image of the MDP $\langle S, A, \Psi, P, R_O \rangle$, with the blocks of the induced partition of Ψ being mostly singletons. The one block that is not a singleton contains all the states not in S'. We can apply our minimization methods reduce the option MDP further. This allows us to abstract away redundancy in the option definition and derive a more compact definition for the option. We refer to this compact option as a *relativized option.* Such options are an extension of the notion of relativized operators introduced by Iba [10]. Formally we define a relativized option as follows:

Definition: A *relativized option* of an MDP $\mathcal{M} = \langle S, A, \Psi, P, R \rangle$ is the tuple $O = \langle h, \mathcal{M}_O, \mathcal{I}, \beta \rangle$, where $\mathcal{I} \subseteq S$ is the initiation set, $\beta : S' \rightarrow [0, 1]$ is the termination function and $h = \langle f, \{g_s | s \in S\} \rangle$ is a partial homomorphism from the MDP $\langle S, A, \Psi, P, R_O \rangle$ to the option MDP $\mathcal{M}_O$ with R_O chosen based on the sub-task.

The option MDP is defined as $\mathcal{M}_O = \langle S' \cup \{\tau\}, A' \cup \{\alpha\}, \Psi', P', R' \rangle$, where $S' = f(S_O)$, where $S_O \subseteq S$ is the domain of O, $\Psi' = h(\Psi)$, P' satisfies conditions (3) and (4) and R' satisfies condition (5). The option policy $\pi : \Psi' \rightarrow [0, 1]$ is obtained by solving $\mathcal{M}_O$ by treating it as an episodic task as before. Note that the initiation set is defined over the state space of $\mathcal{M}$ and not that of $\mathcal{M}_O$. Since the initiation set is typically used by the higher level when invoking the option, we decided to define it over S. When lifted to $\mathcal{M}$, π is suitably transformed into policy fragments over Ψ depending on the state of $\mathcal{M}$ the system is currently in.

Going back to our example in Figure 2(a) we can now define a single relativized option using the option MDP of Figure 2(b) that represents a option to collect the object and leave a room. The policy learned in this option MDP can then be suitable lifted to $\mathcal{M}$ to provide different policy fragments in the different rooms.

4.1 Illustrative Example

We now provide a complete description of the simple gridworld task in Figure 2(a) and some experimental results to illustrate the utility of relativized options. The agent's goal is to collect all the objects in the various rooms by occupying

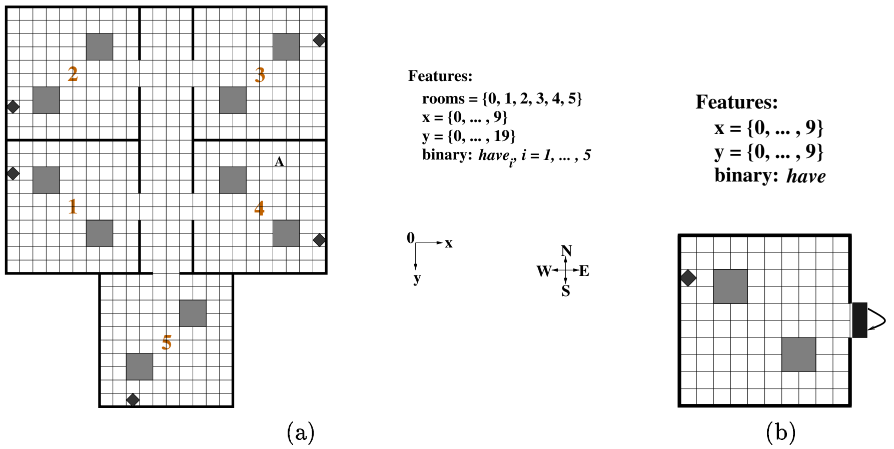


Fig. 2. (a) A simple rooms domain with similar rooms. The task is to collect all 5 objects in the environment. (b) The option MDP corresponding to a *get-object-and-leave-room* option. See text for full description.

the same square as the object. Each of the rooms is a 10 by 10 grid with certain obstacles in it. The actions available to the agent are $\{N, S, E, W\}$ with a 0.1 probability of *failing*, i.e., going randomly in a direction other than the intended one. The state is described by the following features: the room number the agent is in, with 0 denoting the corridor, the x and y co-ordinates within the room or corridor with respect to the reference direction indicated in the figure and boolean variables $have_i$, $i = 0, 1, \ldots, 5$, indicating possession of object in room i. Thus the state with the agent in the cell marked A in the figure and having already gathered the objects in rooms 2 and 4 is represented by $\langle 3, 6, 8, 0, 1, 0, 1, 0 \rangle$. The goal is any state of the form $\langle \cdot, \cdot, \cdot, 1, 1, 1, 1, 1 \rangle$ and the agent receives a reward of $+1$ on reaching a goal state.

We compared the performance of an agent that employs relativized options with that of an agent that uses multiple regular options. The "relativized" agent employs a single relativized option whose policy can be suitably lifted to apply in each of the 5 rooms. The relativized option MDP corresponds to a single room and is shown in Figure 2(b). The state space S' of the option MDP is defined by 3 features: x and y co-ordinates and a binary feature *have*, which is true if the agent has gathered the object in the room. There is an additional absorbing state-action pair (τ, α), otherwise the action set remains the same. The stopping criterion β is 1 at τ and zero elsewhere. The initiation set consists of all states of the form $\langle i, * \rangle$, with $i \neq 0$. There is a reward of $+1$ on transiting to τ from any state of the form $\langle *, 1 \rangle$, i.e. on exiting the room with the object.

One can see that lifting a policy defined in the option MDP yields different policy fragments depending on the room in which the option is invoked. For example, a policy in the option MDP that picks E in all states would lift to yield a policy fragment that picks W in rooms 3 and 4, picks N in room 5 and picks E in rooms 1 and 2.

The "regular" agent employs 5 regular options, $O_1, \cdots, O_5$, one for each room. Each of the option employs the same state space and stopping criterion as the relativized option. The initiation set for option O_i consists of states of the form $\langle i, * \rangle$. There is a reward of +1 on exiting the room with the object. Both agents employ SMDP Q-learning [3] at the higher level and Q-learning [18] at the option level.

We also compared the performance of an agent that employs only the four primitive actions. All the agents used a discount rate of 0.9, learning rate of 0.05 and ϵ-greedy exploration, with an ϵ of 0.1. The results shown are averaged over 100 independent runs. The trials were terminated either on completion of the task or after 3000 steps.

Figure 3(a) shows the asymptotic performance of the agents. This graph demonstrates that the option agents perform similarly in the long run, with no significant difference in performance. The agent that employs only primitive actions takes a long time to start learning and was still improving after 50,000 steps. Since we are more interested in the initial performance of the option agents, we do not present further results for the primitive action agent.

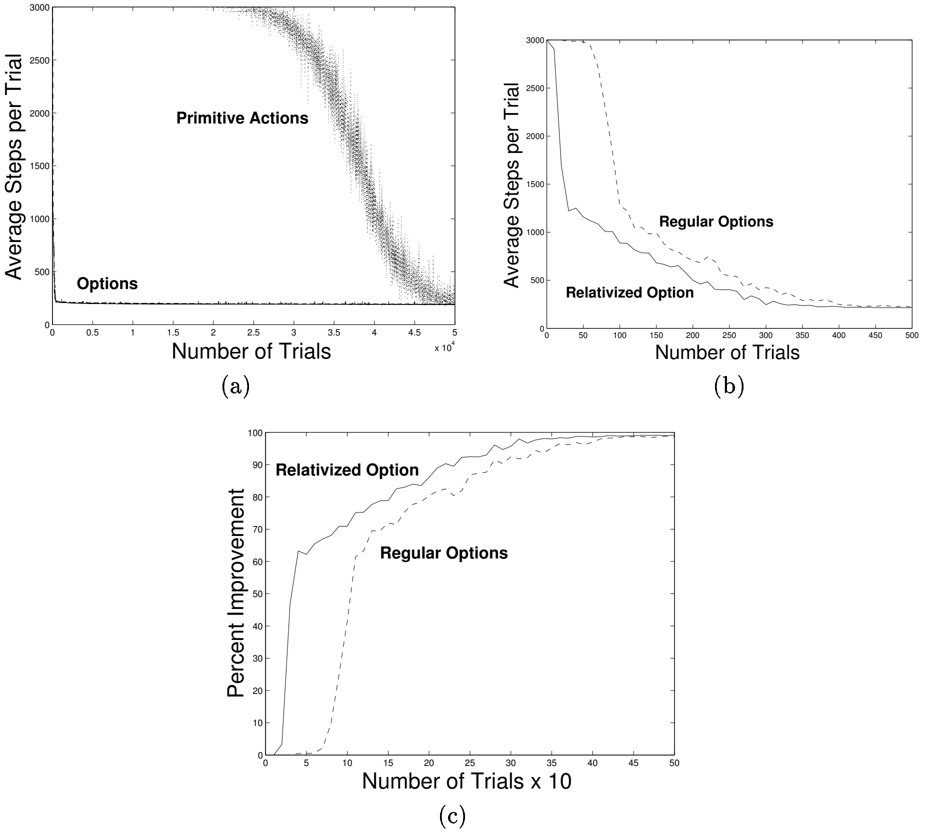
Figure 3(b) shows the initial performance of the option agents. As expected, the relativized agent significantly outperforms the regular agent in the early trials[2]. Figure 3(c) graphs the rate at which the agents improved over their initial performance. The relativized agent achieved similar levels of improvement in performance significantly earlier than the regular option. For example, the relativized agent achieved a 60% improvement in initial performance in 40 trials, while the regular agent needed 110 trials. These results demonstrate that employing relativized options significantly speeds up initial learning performance, and if the homomorphism conditions hold exactly, there is no loss in the asymptotic performance.

5 Approximate Homomorphisms

The various rooms in the test bed above map exactly onto the option MDP in Figure 2(b). In practice such exact equivalences do not arise often. To study the usefulness of relativized options in inexact settings, we conducted further experiments in which the rooms had different dynamics. In the first task, the rooms had the same set of obstacles, but had different probabilities of action success. In the corridor actions fail with probability 0.1 and in rooms 1 through 5 with probabilities 0.2, 0.3, 0.25, 0.5 and 0.0 respectively. Figure 4(b) shows the initial performance of the relativized agent and the regular agent on this task. Again the relativized agent significantly outperforms the regular agent initially and the asymptotic performance, Figure 4(a), shows no significant difference.

In the second task, the rooms have differently shaped obstacles, as shown in Figure 5(a). Again there is a significant improvement in initial performance, but

[2] All the significance tests were *two sample t-tests* with a p-value of 0.01.



(a)

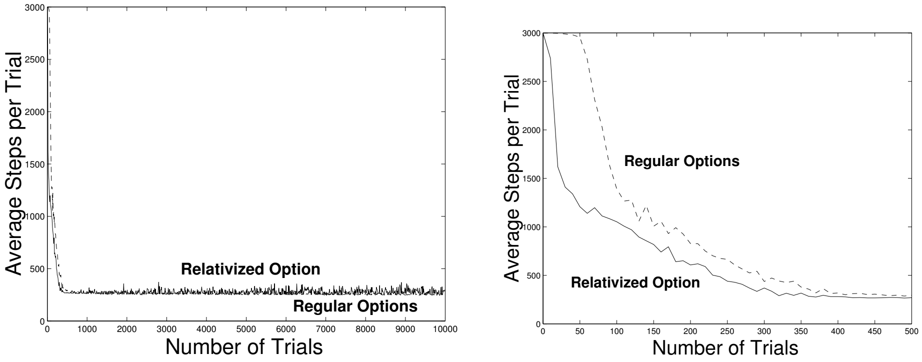
(b)

(c)

Fig. 3. (a) Comparison of asymptotic performance of various learning agents on the task shown in Figure 2. See text for description of the agents. (b) Comparison of initial performance of the regular and relativized agents on the same task. (c) Comparison of the rate of improvement to final performance of the two agents.

the asymptotic performance of the relativized agent is slightly, but significantly, worse than the regular agent, as shown in Figures 6(a) and 6(b). This loss in asymptotic performance is expected and is observed in other inexact scenarios we tested the agents on. In some cases this loss reaches unacceptable levels, with the relativized agent failing to successfully complete the task on certain trials even after considerable training.

One way to bound this loss in asymptotic performance is to model the option homomorphism as a map from an MDP to a *Bounded-parameter MDP* (BMDP) [7]. A BMDP is an MDP in which the transition probabilities and the rewards are specified as intervals. Formally a BMDP $\mathcal{M}'$ is given by the tuple $\langle S, A, \Psi, P_{\updownarrow}, R_{\updownarrow} \rangle$ where S and A are the state and action sets, Ψ is the set of admissible state-action pairs, $P_{\updownarrow} : \Psi \times S \rightarrow [0,1] \times [0,1]$ with $P_{\updownarrow}(s,a,s') = [P_{low}(s,a,s'), P_{high}(s,a,s')]$,



(a) (b)

Fig. 4. (a) Comparison of asymptotic performance of the regular and relativized agents on the modified rooms task. See text for description of the task. (b) Comparison of initial performance of the two agents on the same task.

for all (s, a) in Ψ and s' in S, is the range of values for the probability of transiting from s to s' under action a and $R_{\updownarrow} : \Psi \to \mathbb{R} \times \mathbb{R}$, with $R_{\updownarrow}(s, a) = [R_{low}(s, a), R_{high}(s, a)]$, for all (s, a) in Ψ, is the range of the expected reward on performing action a in state s.

Definition: An *approximate MDP homomorphism h* from an MDP $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ to a BMDP $\mathcal{M}' = \langle S', A', \Psi', P'_{\updownarrow}, R'_{\updownarrow}\rangle$ 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, $\forall s, s' \in S$ and $a \in A_s$:

$$P'_{\updownarrow}(f(s), g_s(a), f(s')) = \left[\min_{t \in [s]_{B_h|S}} T(t, a, [s']_{B_h|S}), \max_{t \in [s]_{B_h|S}} T(t, a, [s']_{B_h|S})\right] \quad (6)$$

$$R'_{\updownarrow}(f(s), g_s(a)) = \left[\min_{t \in [s]_{B_h|S}} R(t, a), \max_{t \in [s]_{B_h|S}} R(t, a)\right] \quad (7)$$

In the rooms task depicted in Figure 5(a) the homomorphism corresponding to the relativized option may now be viewed as a map from each of the rooms to the image BMDP in Figure 5(b), where the probabilities of transiting into and out of the lightly colored states range from 0 to 1. We can now use the *interval value iteration* algorithm of Givan, Leach and Dean [7] to arrive at bounds for the optimal value function in this BMDP and hence can bound the loss of performance that arises due to employing such approximate homomorphisms.

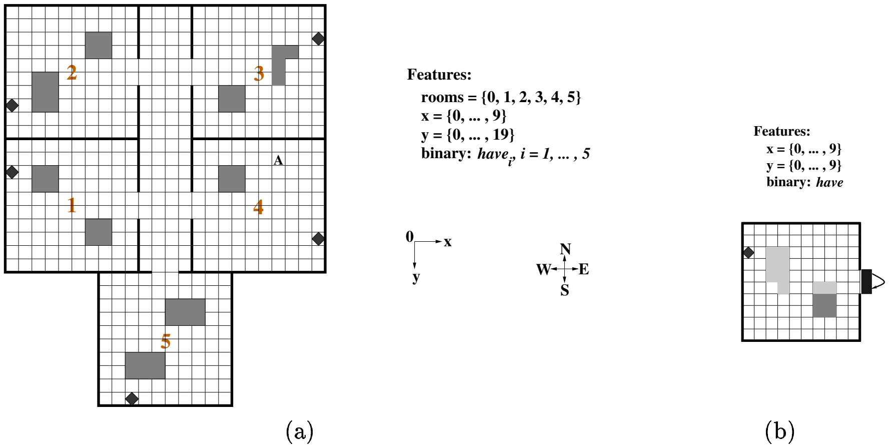


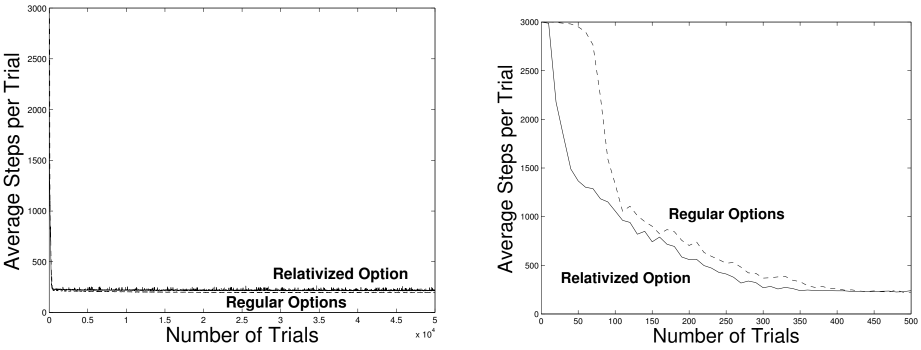
Fig. 5. (a) A simple rooms domain with dissimilar rooms. The task is to collect all 5 objects in the environment. (b) The option BMDP corresponding to a *get-object-and-leave-room* option. See text for full description.

6 Discussion

Our work derives mainly from the model-minimization framework of Dean and Givan [4, 6]. Their work is based on concepts from FSA minimization [9] and concurrent process model checking [13]. They build their framework on the notion of stochastic bisimulations [12] from model checking and extend the definition to MDPs by incorporating the possibility of decision making and stochasticity. But they do not address the problem of state-action equivalence and symmetries. We base our work on the concept of stochastic homomorphisms derived from the FSA literature, which we believe is a simpler notion than bisimulations and helps to better understand the minimization process. Many of the results we present in this paper are extensions of similar results obtained by Givan, Dean and Greig [6] and have counterparts in FSA minimization frameworks.

Minimization algorithms for other modeling paradigms often employ symmetry groups. For example, Jump [11] uses symmetry groups of FSA to decompose a machine into identical components, Emerson and Sistla [5] use symmetry groups to simplify models of concurrent systems, and Glover [8] employs symmetry groups in deriving shift invariant models of Markov processes.

Defining MDP symmetry groups with automorphisms on the states does not capture all the interesting cases of symmetry. Hence employing symmetry groups in Dean and Givan's framework does not give us much leverage. Extending the notion of automorphisms to state-action pairs enables us to overcome this deficiency and employ concepts from group theory and traditional minimization approaches to greater effect. To the best of our knowledge, ours is the first work to employ extended stochastic homomorphisms and symmetry groups in minimization of MDPs.



(a) (b)

Fig. 6. (a) Comparison of asymptotic performance of the regular and relativized agents on the task in Figure 5. (b) Comparison of initial performance of the two agents on the same task.

The state-of-the-art MDP minimization algorithms [1, 2, 6] can automatically construct reduced models of an MDP given the complete system model, i.e., a complete specification of all the components of the MDP. These algorithms can be extended to incorporate state-action equivalence and to compute reduced models as defined in Section 3. Symmetries of MDPs often have special forms. For example Zinkevich and Balch [19] explore symmetries in multi agent systems that arise from permutation of the features corresponding to various agents. Concurrent process literature [5] also abounds with examples of systems with permutation symmetry groups. Often this special form of the symmetry group leads to more efficient minimization algorithms, and we are presently investigating minimization methods that take advantage of symmetries.

Most minimization algorithms, for MDPs and other formalisms, require that we specify the complete system model. Algorithms that exploit symmetries (e.g. ref. [5]) require that we specify the symmetry group beforehand. This requires the designer to provide considerable domain knowledge to the agent, which might not be available or difficult to obtain in many cases. We are currently investigating minimization algorithms that can work with partial specification of the system model and symmetry groups and still derive reasonable reduced models of the system.

Relativized options, per se, do not necessarily add more expressive power to the options framework. It is possible to achieve the same decomposition of a problem by employing regular options. But if we are learning the option policy online, then we garner considerable advantage in terms of efficiency and speed. When employing a relativized option we can employ the experience generated by every invocation of the option to learn a policy on a much smaller image MDP. Thus relativized options allow us to considerably speed up learning and make more efficient use of online experience. They also give us the power to specify a single option that can be applied to many symmetrically equivalent situations, as in the task in Figure 2, where the various rooms are symmetrically equivalent.

In this work we demonstrated that the predicted speed up when employing relativized options is achieved in practice. We also demonstrated that relativized options are useful even in cases where the homomorphism conditions are not satisfied exactly. We employ Bounded-parameter MDPs [7] to characterize approximate homomorphisms and to bound the loss in performance when the homomorphism conditions are not met exactly. The bound on the loss can also be used in establishing conditions under which an option might be relativized and to guide the search for a suitable homomorphism. Our current research is focussed on defining principled ways to generate relativized options.

7 Conclusion

In this paper we presented an MDP minimization framework based on the notion of MDP homomorphism. This is an extension of Dean and Givan's model minimization framework. Our framework can accommodate state-action equivalence and explicitly addresses the issue of modeling symmetries of MDPs. We then developed the concept of partial homomorphisms and applied our minimization framework to hierarchical reinforcement learning to define relativized options—compact, symmetry invariant options. We empirically demonstrated the usefulness of relativized options even in case the homomorphism conditions are not met exactly. We developed the notion of approximate homomorphisms that allow us to bound the loss of performance in such cases.

Acknowledgements

We wish to thank Dan Bernstein, Amy McGovern and Mike Rosenstein for many hours of useful discussion. This material is based upon work supported by the National Science Foundation under Grant No. ECS-9980062 to Andrew G. Barto. 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

1. C. Boutilier and R. Dearden. Using abstractions for decision theoretic planning with time constraints. In *Proceedings of the AAAI-94*, pages 1016–1022. AAAI, 1994.
2. C. Boutilier, R. Dearden, and M. Goldszmidt. Exploiting structure in policy construction. In *Proceedings of International Joint Conference on Artificial Intelligence 14*, pages 1104–1111, 1995.
3. Steven J. Bradtke and Michael O. Duff. Reinforcement learning methods for continuous-time Markov decision problems. In *Advances in Neural Information Processing Systems 7*. MIT Press, 1995.

4. Thomas Dean and Robert Givan. Model minimization in markov decision processes. In *Proceedings of AAAI-97*, pages 106–111. AAAI, 1997.
5. E. A. Emerson and A. P. Sistla. Symmetry and model checking. *Formal Methods in System Design*, 9(1/2):105–131, 1996.
6. Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. Submitted to Artificial Intelligence, 2001.
7. Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter markov decision processes. *Artificial Intelligence*, 122:71–109, 2000.
8. J. Glover. Symmetry groups and translation invariant representations of markov processes. *The Annals of Probability*, 19(2):562–586, 1991.
9. J. Hartmanis and R. E. Stearns. *Algebraic Structure Theory of Sequential Machines.* Prentice-Hall, Englewood Cliffs, NJ, 1966.
10. Glenn A. Iba. A heuristic approach to the discovery of macro-operators. *Machine Learning*, 3:285–317, 1989.
11. J. R. Jump. A note on the iterative decomposition of finite automata. *Information and Control*, 15:424–435, 1969.
12. K. G. Larsen and A. Skou. Bisimulation through probabilistic testing. *Information and Computation*, 94(1):1–28, 1991.
13. D. Lee and M. Yannakakis. Online minimization of transition systems. In *Proceedings of* 24th *Annual ACM Symposium on the Theory of Computing*, pages 264–274. ACM, 1992.
14. Doina Precup. *Temporal Abstraction in Reinforcement Learning.* PhD thesis, University of Massachusetts, Amherst, May 2000.
15. B. Ravindran and A. G. Barto. Symmetries and model minimization of markov decision processes. Technical Report 01-43, University of Massachusetts, Amherst, 2001.
16. Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning. An Introduction.* MIT Press, Cambridge, MA, 1998.
17. Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and Semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.
18. C. J. C. H. Watkins. *Learning from delayed rewards.* PhD thesis, Cambridge University, Cambridge, England, 1989.
19. M. Zinkevich and T. Balch. Symmetry in markov decision processes and its implications for single agent and multi agent learning. In *Proceedings of the 18th International Conference on Machine Learning*, pages 632–640, San Francisco, CA, 2001. Morgan Kaufmann.