An Algebraic Approach to Abstraction in Reinforcement Learning

Balaraman Ravindran and Andrew G. Barto
Department of Computer Science
University of Massachusetts
Amherst, MA, U. S. A.
{ravi|barto}@cs.umass.edu

Abstract

To operate effectively in complex environments learning agents have to selectively ignore irrelevant details by forming useful abstractions. In this article we outline a formulation of abstraction for reinforcement learning approaches to stochastic sequential decision problems modeled as semi-Markov Decision Processes (SMDPs). Building on existing algebraic approaches, we propose the concept of SMDP homomorphism and argue that it provides a useful tool for a rigorous study of abstraction for SMDPs. We apply this framework to different classes of abstractions that arise in hierarchical systems and discuss *relativized options*, a framework for compactly specifying a related family of temporally-extended actions. Additional details of this work are described in refs. [1, 2, 3].

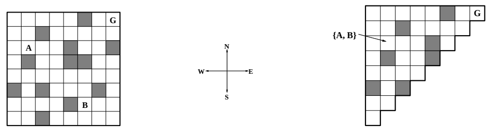
1 Introduction

The ability to form abstractions is one of the features that allows humans to operate effectively in complex environments. We systematically ignore information that we do not need for performing an immediate task at hand. While driving, for example, we can ignore details regarding our clothing and the state of our hair. Researchers in artificial intelligence (AI), in particular machine learning (ML), have long recognized that applying computational approaches to operating and learning in complex and real-world environments requires the ability to form and manipulate useful abstractions. In this article we outline elements of an algebraic approach to abstraction that builds on early research on the algebraic theory of abstract automata, adapting it to stochastic sequential decision problems modeled as Markov decision processes (MDPs) and semi-Markov decision processes (SMDPs). The latter formalism is widely used in recent approaches to extending reinforcement learning (RL) methods to hierarchical systems [4, 5, 6].

We introduce the concept of an *SMDP homomorphism* and argue that it provides a unified view of key issues essential for a rigorous treatment of abstraction for stochastic dynamic decision processes. The concept of a homomorphism between dynamic systems, sometimes called a “dynamorphism” [7], has played an important role in theories of abstract automata [8], theories of modeling and simulation [9], and is frequently used by researchers studying model checking approaches to system validation [10]. Although those studying approximation and abstraction methods for MDPs and SMDPs have employed formalisms that implicitly embody the idea of a homomorphism, they have not made explicit use of the appropriate homomorphism concept. We provide what we claim is the appropriate concept and give examples of how it can be widely useful as the basis of abstraction in stochastic dynamic settings. Additional details of this work are described in refs. [1, 2, 3].

Informally, the kind of homomorphism we consider is a mapping from one dynamic system to another that eliminates state distinctions while preserving the system’s dynamics. We present a definition of homomorphism that is appropriate for SMDPs. In ref. [2] we developed an MDP abstraction framework based on MDP homomorphisms. This extended the MDP minimization framework proposed by Dean and Givan [11] and enabled the accommodation of redundancies arising from symmetric equivalence of the kind illustrated in Figure 1.

We then extend the notion of SMDP homomorphism to hierarchical systems. In particular, we apply homomorphisms in the *options* framework introduced by Sutton, Precup and Singh [4] to pro-



(a) (b)

Figure 1: (a) A symmetric gridworld problem. The goal state is G and there are four deterministic actions. This gridworld is symmetric about the NE-SW diagonal. For example, states A and B are equivalent since for each action in A, there is an equivalent action in B. Taking action E, say, in state A is equivalent to taking action N in state B, in the sense that they go to equivalent states that are each one step closer to the goal. (b) An equivalent 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.

vide a formal basis for planning and learning with temporally-extended actions. We argue that this use of the SMDP homomorphism concept facilitates employing different abstractions at different levels of a hierarchy. We also discuss *relativized options*, a framework for defining "option schema". Here an option is defined in a relative frame of reference and can be transformed to suit a particular situation when it is invoked.

After introducing some notation (Section 2), we define SMDP homomorphisms and discuss modeling symmetries (Section 3). Then we discuss our approach to abstraction in hierarchical systems (Section 4) and conclude with some discussion of directions for future research (Section 5).

2 Notation

A (finite) *Markov Decision Process* is a tuple $\langle S, A, \Psi, P, R\rangle$, where $S = \{1, 2, \cdots, n\}$ is a 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 discrete time semi-Markov decision process (SMDP) is a generalization of an MDP in which actions can take variable amounts of time to complete. As with an MDP, an SMDP is a tuple $\langle S, A, \Psi, P, R\rangle$, where S, A and Ψ are the sets of states, actions and admissible state-action pairs; $P : \Psi \times S \times \mathbb{N} \to [0, 1]$ is the transition probability function with $P(s, a, s', N)$ being the probability of transition from state s to state s' under action a in N time steps, and $R : \Psi \times \mathbb{N} \to \mathbb{R}$ is the expected discounted reward function, with $R(s, a, N)$ being the expected reward for performing action a in state s and completing it in N time steps.[1]

A (stationary) *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 executing action a in state s. The *value* of a state-action pair (s, a) under policy π is the expected value of the sum of discounted future rewards starting from state s, taking action a, and following π thereafter. When the SMDP has well defined terminal states, we often do not discount future rewards. In such cases an SMDP is equivalent to an MDP and we will ignore the transition times. The *action-value function*, Q^π, corresponding to a policy π is the mapping from state-action pairs to their values. 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 (or equivalence relation) on X, with $[x]_f = [x']_f$ if and only if $f(x) = f(x')$.

3 SMDP 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 state that it preserves [9]. The specific definition of homomorphism that we claim is most useful for MDPs and SMDPs is as follows:

Definition: An *SMDP homomorphism* h from an SMDP $\mathcal{M} = \langle S, A, \Psi, P, R\rangle$ to an SMDP $\mathcal{M}' = \langle S', A', \Psi', P', R'\rangle$ is a surjection from Ψ to Ψ', defined by a tuple of surjections $\langle f, g_1, g_2, \cdots, g_n\rangle$, 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 $\forall s, s' \in$

[1] We are adopting the formalism of Dietterich [5].

$S, a \in A_s$ and for all $N \in \mathbb{N}$:

$$P'(f(s), g_s(a), f(s'), N) = \sum_{t \in [s']_f} P(s, a, t, N) \tag{1}$$

$$R'(f(s), g_s(a), N) = R(s, a, N). \tag{2}$$

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 SMDP $\mathcal{M}'$ 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 [8]. Condition (2) says that state-action pairs that have the same image under h have the same expected reward. An MDP homomorphism is similar to an SMDP homomorphism except that the conditions (1) and (2) apply only to the states and actions and not to the transition times.

The state-dependent action mapping allows us to model symmetric equivalence in MDPs and SMDPs. For example, if $h = \langle f, g_1, g_2, \cdots, g_n \rangle$ 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. Whereas Zinkevich and Balch [12] defined symmetries of MDPs by employing equivalence relations on the state-action pairs, we explicitly formalize the notion of SMDP symmetries employing SMDP homomorphisms and group theoretic concepts.

Definitions: An SMDP homomorphism $h = \langle f, g_1, g_2, \cdots, g_n \rangle$ from SMDP $\mathcal{M} = \langle S, A, \Psi, P, R \rangle$ to SMDP $\mathcal{M}' = \langle S', A', \Psi', P', R' \rangle$ is an *SMDP 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 SMDP isomorphism from an SMDP $\mathcal{M}$ to itself is an *automorphism* of $\mathcal{M}$.

The set of all automorphisms of an SMDP $\mathcal{M}$, denoted by Aut$\mathcal{M}$, forms a group under composition of homomorphisms. This group is the *symmetry group* of $\mathcal{M}$. In the gridworld example of Figure 1, the symmetry group consists of the identity map on states and actions, a reflection of the states about the NE-SW diagonal and a swapping of actions N and E and of actions S and W. Any subgroup of the symmetry group of an SMDP induces an equivalence relation on Ψ, which can also be induced by a suitably defined homomorphism [1]. Therefore we can model symmetric equivalence as a special case of homomorphic equivalence.

The notion of homomorphic equivalence immediately gives us an SMDP minimization framework. In ref. [1] we extended the minimization framework of Dean and Givan [11, 13] 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. We also showed that when $\mathcal{M}'$ is a homomorphic image of an MDP $\mathcal{M}$, 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 induce an optimal policy in $\mathcal{M}$. Thus we can solve the original MDP by solving a homomorphic image. It is easy to extend these results to SMDP models.

While we can derive reduced models with a smaller state set by applying minimization ideas, we do not necessarily simplify the description of the problem in terms of the number of parameters required. But MDPs often have additional structure associated with them that can be exploited to develop compact representations. By specializing the definition of SMDP homomorphism to systems whose states are vectors of values of descriptive variables, we can model abstraction schemes for structured MDPs. In ref. [3] we present a simple example of such an abstraction scheme that employs simple structured homomorphisms. Without suitable constraints, often derived from prior knowledge of the structure of the problem, searching for general structured homomorphisms results in a combinatorial explosion. Abstraction algorithms developed by Boutilier and colleagues can be modeled as converging to constrained forms of structured morphisms assuming various representations of the conditional probability tables—when the space of morphisms is defined by Boolean formulae of the features [14], when it is defined by decision trees on the features [15], and when it is defined by first-order logic formulae [16].

4 Abstraction in Hierarchical Systems

SMDP homomorphisms can readily be employed to model various abstraction schemes in "flat" MDPs and SMDPs. SMDP homomorphisms are a convenient and powerful formalism for modeling abstraction schemes in hierarchical systems as well. Before describing various abstraction approaches, we first introduce a hierarchical architecture that supports abstraction.

4.1 Hierarchical Markov Options

Recently several hierarchical reinforcement learning frameworks have been proposed [6, 4, 5] all of which use the SMDP formalism. In this article the hierarchical framework we adopt is the *options* framework [4], although the ideas developed here are more generally applicable. Options are actions that take multiple time steps to complete. They are usually described by the following components: the policy the agent follows while the option is executing, the set of states in which the option can begin execution, and a termination function, $\beta : S \rightarrow [0, 1]$, which gives the probability with which the option can terminate in each state. The resulting system is naturally modeled as an SMDP with the transition time distributions induced by the option policies. We present an extension to the options framework that readily facilitates modeling abstraction at multiple levels of the hierarchy using SMDP homomorphisms.

We consider the class of options known as Markov options, whose policies satisfy the Markov property and that terminate on achieving a certain sub-goal. In such instances it is possible to implicitly define the option policy as the solution to an *option MDP*, or an option SMDP if the option has access to other options, that is, if its policy can “call” other options. Accordingly we have the following definition:

Definition: A *hierarchical Markov sub-goal option* of an SMDP $\mathcal{M} = \langle S, A, \Psi, P, R \rangle$ is the tuple $O = \langle \mathcal{M}_O, \mathcal{I}, \beta \rangle$, where $\mathcal{I} \subseteq S$ is the initiation set of the option, $\beta : S \rightarrow [0, 1]$, is the termination function and $\mathcal{M}_O$ is the option SMDP.

The state set of $\mathcal{M}_O$ is a subset of S and constitutes the *domain* of the option. The action set of M_O is a subset of A and may contain other options as well as “primitive” actions in A. The reward function of $\mathcal{M}_O$ is chosen to reflect the sub-goal of the option. The transition probabilities of $\mathcal{M}_O$ are induced by P and the policies of lower level options. We assume that the lower-level options are following fixed policies which are optimal in the corresponding option SMDPs. The option policy π is obtained by solving $\mathcal{M}_O$, treating it as an episodic task 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 an example, refer to the simple gridworld task shown in Figure 2(a). Here, an option to pick up the object and exit room 1 can be defined as the solution to the problem shown in 2(b), with a suitably defined reward function. The domain and the initiation set of the option consists of all the states in the room, and the option terminates when the agent exits the room with or without the object.

To learn with hierarchical Markov options we may employ hierarchical SMDP Q-learning [17, 18], where the lowest levels of the hierarchy use Q-learning and the higher levels use SMDP Q-learning. In earlier work we showed empirically that simultaneously learning at multiple levels of the hierarchy converges to a recursively optimal solution, i.e., a solution that is optimal given that all the lower level solutions are recursively optimal. In fact, it can be shown that under the usual assumptions on the learning rate and exploration policy, hierarchical SMDP Q-learning with suitably defined hierarchal Markov options always converges to a recursively optimal policy, even when learning simultaneously at all levels of the hierarchy. The proof of this statement follows along the lines of Dietterich [5].

4.2 Option Specific Abstraction

The homomorphism conditions (1) and (2) are very strict and frequently we end up with trivial homomorphic images when deriving abstractions based on a non-hierarchical SMDP. But it is often possible to derive non-trivial reductions if we restrict attention to certain sub-problems, i.e., certain sub-goal options. In such cases we can apply the ideas discussed in Section 3 to an option SMDP directly to derive abstractions that are specific to that option. The problem of learning the option policy is transformed to the usually simpler problem of learning an optimal policy for the homomorphic image.

Dietterich [5] introduced safe state-abstraction conditions for the MaxQ architecture, a hierarchical learning framework related to the options framework. These conditions ensure that the resulting abstractions do not result in any loss of performance. He assumes that the sub-problems at different levels of the hierarchy are specified by factored MDPs. In ref. [3] we show that the homomorphism conditions are a generalization of Dietterich's abstraction conditions as applicable to the hierarchical Markov options framework.

4.3 Relativized Options

In this section we explore in more detail one of the implications of employing homomorphic images as option MDPs. Consider the problem of navigating in the gridworld environment shown in Figure 2(a). The goal is to reach the central corridor after collecting all the objects in the environment. No non-trivial homomorphic image exists of the entire problem. But there are many similar components in the problem, namely, the five sub-tasks of getting the object and exiting $room_i$.

We can model these similar components by a "partial" homomorphic image—where the homomorphism conditions are applicable only to states in a given room. One such partial image is shown in Figure 2(b). Employing such an abstraction lets us compactly represent a related family of options, in this case the tasks of collecting objects and exiting each of the five rooms, using a single option MDP. We refer to this compact option as a *relativized option.* Such abstractions are an extension of the notion of relativized operators introduced by Iba [19]. Formally we define a relativized option as follows:

Definition: A *relativized option* of an SMDP $\mathcal{M} = \langle S, A, \Psi, P, R \rangle$ 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 $h = \langle f, g_1, g_2, \cdots, g_n \rangle$ is a partial homomorphism from the SMDP $\langle S, A, \Psi, P, R_O \rangle$ to the option SMDP $\mathcal{M}_O$ with R_O chosen based on the sub-task.

Here the state set of $\mathcal{M}_O$ is $S' = f(S_O)$, where S_O is the domain of the option, and the admissible state-action set is $h(\Psi)$. Going back to the example in Figure 2(a), we can now define a single *get-object-and-leave-room* relativized option using the option MDP of Figure 2(b). The policy learned in this option MDP can then be suitably lifted to $\mathcal{M}$ to provide different policy fragments in the different rooms. Figure 3 demonstrates the speed-up in learning when using a single relativized option as opposed to five regular options. In this experiment the option policies and the higher level policy were learned simultaneously. In ref. [2] we have reported more detailed experiments in this setting.

5 Discussion

The equivalence classes induced by SMDP homomorphisms satisfy the stochastic version of the substitution property [8]. This property is also closely related to *lumpability* in Markov chains [20] and *bisimulation homogeneity* [13] in MDPs. We chose the SMDP homomorphism as our basic formalism because we believe that it is a simpler notion and provides a more intuitive explanation of various abstraction schemes.

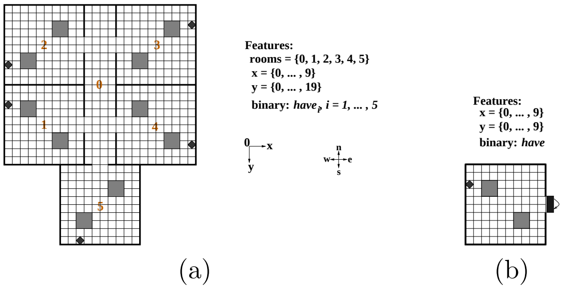


Figure 2: (a) A simple rooms domain with similar rooms and usual stochastic gridworld dynamics. The task is to collect all 5 objects (black diamonds) in the environment and reach the central corridor. The shaded squares are obstacles. (b) The option MDP corresponding to a *get-object-and-leave-room* option.

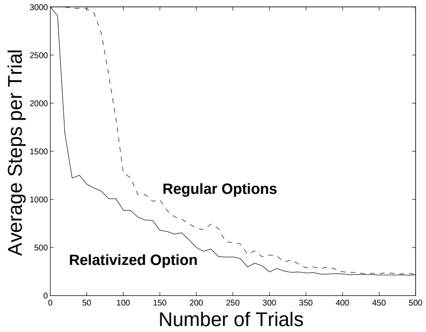


Figure 3: Comparison of performance of learning agents employing regular and relativized options on the task shown in Figure 2.

The homomorphism conditions (1) and (2) are very strict conditions that are often not met exactly in practice. One approach is to relax the homomorphism conditions somewhat and allow small variations in the block transition probabilities and rewards. We have explored this issue in ref. [2], basing our approximate homomorphisms on the concept of *Bounded-parameter MDPs* developed by Givan, Leach and Dean [21]. We are currently working on extending approximate homomorphisms to hierarchical systems so as to accommodate variations in transition-time distributions.

Although SMDP homomorphisms are powerful tools for modeling abstraction, finding a minimal image of a given SMDP is an NP-hard problem. While taking advantage of structure allows us to develop efficient algorithms in special cases, much work needs to be done to develop efficient general purpose algorithms. Currently we are investigating methods that allow us to determine homomorphisms given a set of candidate transformations in a hierarchical setting.

In this article we described a novel definition of SMDP homomorphism that employs state-dependent recoding of actions. This allows us to extend existing minimization and abstraction methods to a richer class of problems. We then described how this formulation of abstraction can be useful in the construction of hierarchical learning architectures. We believe that SMDP homomorphism can serve as the basis for modeling a variety of abstraction paradigms.

Acknowledgments

We thank Dan Bernstein and Mike Rosenstein for many hours of useful discussion and Bob Givan and Matt Greig for clarifying certain ideas from their work. This material is based on 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

[1] B. Ravindran and A. G. Barto. Symmetries and model minimization of Markov decision processes. Technical Report 01-43, University of Massachusetts, Amherst, 2001.

[2] B. Ravindran and A. G. Barto. Model minimization in hierarchical reinforcement learning. In Sven Koenig and Robert C. Holte, editors, *Proceedings of the Fifth Symposium on Abstraction, Reformulation and Approximation (SARA 2002), Lecture Notes in Artificial Intelligence 2371*, pages 196–211, New York, NY, August 2002. Springer-Verlag.

[3] B. Ravindran and A. G. Barto. Smdp homomorphisms: An algebraic approach to abstraction in semi-Markov decision processes. In *Proceedings of the Eighteenth Internatinal Joint Conference on Artificial Intelligence (IJCAI 2003)*, August 2003.

[4] R. S. Sutton, D. Precup, and S. P. Singh. Between MDPs and Semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.

[5] T. G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Artificial Intelligence Research*, 13:227–303, 2000.

[6] R. Parr and S. Russell. Reinforcement learning with hierarchies of machines. In *Proceedings of Advances in Neural Information Processing Systems 10*, pages 1043–1049. MIT Press, 1997.

[7] M. A. Arbib and E. G. Manes. *Arrows, Structures and Functors*. Academic Press, New York, NY, 1975.

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

[9] B. P. Zeigler. On the formulation of problems in simulation and modelling in the framework of mathematical system theory. In *Proceedings of the Sixth International Congress on Cybernetics*, pages 363–385. Association Internationale de Sybernétique, 1972.

[10] E. A. Emerson and A. P. Sistla. Symmetry and model checking. *Formal Methods in System Design*, 9(1/2):105–131, 1996.

[11] T. Dean and R. Givan. Model minimization in Markov decision processes. In *Proceedings of AAAI-97*, pages 106–111. AAAI, 1997.

[12] 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.

[13] R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in Markov decision processes. To appear in Artificial Intelligence, 2003.

[14] 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.

[15] 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.

[16] C. Boutilier, R. Reiter, and R. Price. Symbolic dynamic programming for first-order mdps. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 541–547, 2001.

[17] S. J. Bradtke and M. O. Duff. Reinforcement learning methods for continuous-time Markov decision problems. In S. Minton, editor, *Advances in Neural Information Processing Systems*, volume 7, pages 393–400, Cambridge, MA., 1995. MIT Press.

[18] T. G. Dietterich. An overview of MAXQ hierarchical reinforcement learning. In B. Y. Choueiry and T. Walsh, editors, *Proceedings of the Fourth Symposium on Abstraction, Reformulation and Approximation SARA 2000, Lecture Notes in Artificial Intelligence*, pages 26–44, New York, NY, 2000. Springer-Verlag.

[19] G. A. Iba. A heuristic approach to the discovery of macro-operators. *Machine Learning*, 3:285–317, 1989.

[20] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Van Nostrand, Princeton, NJ, 1960.

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