Value Preserving State-Action Abstractions

David Abel
Brown University

Nathan Umbanhowar
Brown University

Khimya Khetarpal
Mila-McGill University

Dilip Arumugam
Stanford University

Doina Precup
Mila-McGill University

Michael L. Littman
Brown University

Abstract

Abstraction can improve the sample efficiency of reinforcement learning. However, the process of abstraction inherently discards information, potentially compromising an agent's ability to represent high-value policies. To mitigate this, we here introduce combinations of state abstractions and options that are guaranteed to preserve the representation of near-optimal policies. We first define ϕ-relative options, a general formalism for analyzing the value loss of options paired with a state abstraction, and present necessary and sufficient conditions for ϕ-relative options to preserve near-optimal behavior in any finite Markov Decision Process. We further show that, under appropriate assumptions, ϕ-relative options can be composed to induce hierarchical abstractions that are also guaranteed to represent high-value policies.

1 INTRODUCTION

Learning to make high-quality decisions in complex environments is challenging. To mitigate this difficulty, knowledge can be incorporated into learning algorithms through inductive biases, heuristics, or other priors that provide information about the world. In reinforcement learning (RL), one powerful class of such structures takes the form of *abstractions*, either of state (what are the relevant properties of the world?) or action (what correlated, long-horizon sequences of actions are useful?).

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

To make RL more tractable, abstractions throw away information about the environment such as irrelevant state features or ineffective sequences of actions. If the abstractions are *too* aggressive, however, they destroy an agent's ability to solve tasks of interest. Thus, there is a trade-off inherent in the role that abstractions play: they should 1) make learning easier, while 2) preserving enough information to support the discovery of good behavioral policies. Indeed, prior work has illustrated the potential for abstractions to support both this first (Konidaris and Barto, 2007; Brunskill and Li, 2014; Bacon et al., 2017) and second property (Li et al., 2006; Van Roy, 2006; Hutter, 2014).

To realize the benefits of state *and* action abstraction, it is common to make use of both. One approach builds around *MDP homomorphisms*, introduced by Ravindran and Barto (2002), based on the earlier work of Givan et al. (1997) and Whitt (1978). Certain classes of homomorphisms can lead to dramatic reductions in the size of the model needed to describe the environment while preserving representation of high value policies. Ravindran and Barto (2003a, 2004) extend these ideas to semi-Markovian environments based on the *options* framework (Sutton et al., 1999), illustrating how approximate model reduction techniques can be blended with hierarchical structures to form good compact representation. Majeed and Hutter (2019) carry out similar analysis in non-Markovian environments, again proving which conditions preserve value. However, we lack a general theory of value preserving state-action abstractions, especially when abstract actions are extended over multiple time steps. More concretely, the following question remains open:

Which combinations of state abstractions and options preserve representation of near-optimal policies?

The primary contribution of this work is new analysis clarifying which combinations of state abstractions (ϕ) and options ($\mathcal{O}$) preserve representation of near-optimal policies in finite Markovian environments.

To perform this analysis, we first define ϕ-relative options, a general formalism for analyzing the value loss of a state abstraction paired with a set of options. We then prove four sufficient conditions, along with one necessary condition, for ϕ-relative options to preserve near-optimal behavior in any finite Markov Decision Process (MDP) (Bellman, 1957; Puterman, 2014). We further prove that ϕ-relative options can be composed to induce a hierarchy that preserves near-optimal behavior under appropriate assumptions about the hierarchy's construction. We suggest these results can support the development of principled methods that learn and make use of value-preserving abstractions.

1.1 Background

We next provide background on state and action abstraction. We take the standard treatment of RL: an agent learns to make decisions that maximize value while interacting with an MDP denoted by the five tuple $(\mathcal{S}, \mathcal{A}, R, T, \gamma)$. For more on RL, see the book by Sutton and Barto (2018), for more on MDPs, see the book by Puterman (2014).

A state abstraction is an aggregation function that projects the environmental state space into a smaller one. With a smaller state space, algorithms can learn with less computation, space, and samples (Singh et al., 1995; Dearden and Boutilier, 1997; Dietterich, 2000b,a; Andre and Russell, 2002; Jong and Stone, 2005). However, throwing away information about the state space can restrict an agent's ability to represent good policies; as a trivial illustration of this, consider the maximally-compressing state abstraction that maps all original states in the environment to a single abstract state. Only well-chosen types of state abstraction are known to preserve representation of good policies (Dean and Givan, 1997; Li et al., 2006; Van Roy, 2006; Hutter, 2014; Abel et al., 2016, 2019; Majeed and Hutter, 2019). We define a state abstraction as follows:

Definition 1 (State Abstraction): *A state abstraction $\phi : \mathcal{S} \to \mathcal{S}_\phi$ maps each ground state, $s \in \mathcal{S}$, into an abstract state, $s_\phi \in \mathcal{S}_\phi$.*

Since the abstract state space is *smaller* than the original MDP's state space, we expect to lose information. In the context of sequential decision making, it is natural to permit this information loss so long as RL algorithms can still learn to make useful decisions (Abel et al., 2019; Harutyunyan et al., 2019).

Next, we recall options, a popular formalism for organizing the action space of an agent.

Definition 2 (Option (Sutton et al., 1999)): *An option $o \in \mathcal{O}$ is a triple $(\mathcal{I}_o, \beta_o, \pi_o)$, where $\mathcal{I}_o \subseteq \mathcal{S}$ is a subset of the state space denoting where the option initiates; $\beta_o \subseteq \mathcal{S}$, is a subset of the state space denoting where the option terminates; and $\pi_o : \mathcal{S} \to \Delta(\mathcal{A})$ is a policy prescribed by the option o.*

Options denote abstract actions; the three components indicate where the option o can be executed ($\mathcal{I}_o$), where it terminates (β_o), and what to do in between these two conditions (π_o). Options are known to aid in transfer (Konidaris and Barto, 2007, 2009; Brunskill and Li, 2014; Topin et al., 2015), encourage better exploration (Şimşek and Barto, 2004, 2009; Brunskill and Li, 2014; Bacon et al., 2017; Fruit and Lazaric, 2017; Machado et al., 2017; Tiwari and Thomas, 2019; Jinnai et al., 2019), and make planning more efficient (Mann and Mannor, 2014; Mann et al., 2015). We define an action abstraction in terms of options as follows:

Definition 3 (Action Abstraction): *An action abstraction is any replacement of the primitive actions, $\mathcal{A}$, with a set of options $\mathcal{O}$.*

Thus, an RL algorithm paired with an action abstraction chooses from among the available options at each time step, runs the option until termination, and then chooses the next option. With $\mathcal{O}$ *replacing* the primitive action space,[1] we are no longer guaranteed to be able to represent every policy and may destroy an agent's ability to discover a near-optimal policy. If the primitive actions are included, any policy that can be represented in the original state-action space can still be represented. However, by including options *and* primitive actions, learning algorithms face a larger branching factor, and must search the full space of policies which can hurt learning performance (Jong et al., 2008). Hence, it is often prudent to restrict the action space *only* to a set of options to avoid blowing up the search space. Ideally, we would make use of options that both 1) preserve representation of good policies, while 2) keeping the branching factor and induced policy class small. Naturally, using the options that always execute the optimal policy maximally satisfies both properties, but are challenging to obtain. Establishing a clear understanding of how this first property can be satisfied under approximate knowledge is the primary motivation of this work.

2 RELATED WORK

The study of abstraction in RL has a long and exciting history, dating back to early work on approximating dynamic programs by Fox (1973) and Whitt (1978, 1979), along with the early work on hierar-

[1] In this work, we are uninterested in degenerate options (Bacon et al., 2017; Harb et al., 2018) that alias individual primitive actions ($\mathcal{I}_o = \beta_o = \mathcal{S}$, $\exists a : \pi_o(a \mid s) = 1$).

chical RL (Dayan and Hinton, 1993; Wiering and Schmidhuber, 1997; Parr and Russell, 1998; Dietterich, 2000a), options (Sutton et al., 1999), and state abstraction (Tsitsiklis and Van Roy, 1996; Dean and Givan, 1997; Andre and Russell, 2002; Li et al., 2006). This literature is vast; we here concentrate on work that is focused on abstractions that aim to retain representation of near-optimal behavior. We divide this summary into each of state, action, joint state-action, and hierarchical abstraction.

State Abstraction. The work of Whitt (1978, 1979) paved the way for understanding the value loss of state abstraction in MDPs. Later, Dean and Givan (1997) developed a method for finding states that behave similarly to one another via the *bisimulation* property (Larsen and Skou, 1991). Many subsequent works have explored bisimulation for abstraction, including defining metrics for finite (Ferns et al., 2004; Castro and Precup, 2011; Castro, 2019) and infinite MDPs (Ferns et al., 2006; Taylor et al., 2008). In a similar vein, Li et al. (2006) provide a unifying framework for state abstraction in MDPs, examining when such abstractions will preserve optimal behavior and affect existing convergence guarantees of well-known RL algorithms. Recent work has continued to clarify the conditions under which state abstractions preserve value in MDPs (Jiang et al., 2015a; Abel et al., 2016), non-Markovian environments (Hutter, 2014), planning algorithms (Hostetler et al., 2014; Jiang et al., 2014; Anand et al., 2015) and lifelong RL (Abel et al., 2018). A related and important body of work studies the problem of *selecting* a state abstraction from a given class (Maillard et al., 2013; Odalric-Ambrym et al., 2013; Jiang et al., 2015a; Ortner et al., 2019).

Action Abstraction. Little is known about which kinds of options or action abstractions preserve value; this stems largely from the fact that most prior studies of options consider the setting where options are *added* to the primitive actions, so $\mathcal{A}_{\text{abstract}} = \mathcal{A} \cup \mathcal{O}$. Since the primitive actions are included, the value loss is always zero as the full space of policies is still representable. Brunskill and Li (2014) study the problem of learning options that improve sample efficiency in lifelong RL. Under mild assumptions, their algorithm finds options that let will match some default sample complexity (without options). However, in order to realize the benefits of options, it is important to consider the removal of primitive actions—otherwise, RL algorithms can still represent the full space of policies, and so learning speed may suffer (Jong et al., 2008).

A few previous works study value preservation in the case where options replace the primitive actions. Most recently, Mann and Mannor (2014); Mann et al. (2015) investigate the effect of options on approximate planning algorithms. Their analysis considers options that use ε-optimal policies and are guaranteed to run for a large number of time steps, ensuring that few mistakes are made by each option. The key result of Mann et al. (2015) characterizes the convergence rate of value iteration using these kinds of options, with this rate depending directly on the sub-optimality and length of the options. Lehnert et al. (2018) explore the impact of horizon length on representation of value functions. They find that the value loss of any policy that optimizes with respect to an artificially-short horizon can achieve value similar to that of policies that take into account the full horizon, building on the results of Jiang et al. (2015b). Similar analysis is conducted in the case of well connected subsets of states, paralleling our treatment of state abstractions here.

State-Action Abstraction. Together, state and action abstractions can distill complex problems into simple ones (Jonsson and Barto, 2001; Ciosek and Silver, 2015). One popular approach builds around *MDP homomorphisms* (Ravindran and Barto, 2002, 2003a,b, 2004; Ravindran, 2003). MDP homomorphisms compress the MDP by collapsing state-action pairs that can be treated as equivalent. The recent work by Majeed and Hutter (2019) extends this analysis to non-Markovian settings, proving the existence of several classes of value preserving homomorphisms. The key difference between their results and the analysis we present here is our close attachment to the options formalism. Indeed, as we will show, our framework for state-action abstractions can capture MDP homomorphisms in a sense.

Separately, Bai and Russell (2017) develop a Monte Carlo planning algorithm that incorporates state and action abstractions to efficiently solve the Partially Observable MDP (Kaelbling et al., 1998) induced by the abstractions. Theorem 1 from Bai and Russell (2017) shows that the value loss of their approach is bounded as a function of the state aggregation error (Hostetler et al., 2014), and Theorem 2 shows their algorithm converges to a recursively optimal policy for given state-action abstractions. Our results are closely related, but capture broad classes of state-action abstraction (Theorem 1) and target global optimality, rather than recursive optimality (Theorem 3).

Hierarchical Abstraction. Just as state and action abstractions enable a coarser view of a decision-making problem through a single lens, many approaches accommodate abstraction at multiple levels of granularity (Dayan and Hinton, 1993; Wiering and Schmidhuber, 1997; Parr and Russell, 1998; Dietterich, 2000a; Barto and Mahadevan, 2003; Jong and

Stone, 2008; Konidaris, 2016; Bai and Russell, 2017; Konidaris et al., 2018; Levy et al., 2019). Most recently, Nachum et al. (2019) introduce a scheme for preserving near-optimal behavior in hierarchical abstractions. Their main result studies the kinds of multi-step state visitation distributions induced by different hierarchies. In particular, they show that hierarchies that can induce a sufficiently rich space of k-step state distributions can represent near-optimal behavior, too. Indeed, this represents the strongest existing result for how to construct value-preserving hierarchies based on approximate knowledge.

3 ANALYSIS: ϕ-RELATIVE OPTIONS

We incorporate state and action abstraction into RL as follows. When the environment transitions to a new state s, the agent processes s via ϕ yielding the abstract state, s_ϕ. Then, the agent chooses an option from among those that initiate in s_ϕ, and follow the chosen option's policy until termination, where this process repeats. In this way, an RL agent can reason in terms of abstract state and action alone, without knowing the true state or action space.

To analyze the value loss of these joint abstractions, we first introduce ϕ-relative options, a novel means of combining state abstractions with options.

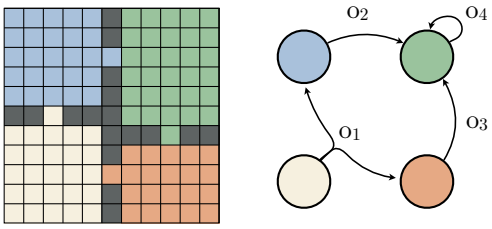
Definition 4 (ϕ-Relative Option): *For a given ϕ, an option is said to be ϕ-relative if and only if there is some $s_\phi \in \mathcal{S}_\phi$ such that, for all $s \in \mathcal{S}$:*

$$\mathcal{I}_o(s) \equiv s \in s_\phi, \qquad \beta_o(s) \equiv s \notin s_\phi, \qquad \pi_o \in \Pi_{s_\phi}, \quad (1)$$

where $\Pi_{s_\phi} : \{s \mid \phi(s) = s_\phi\} \to \Delta(\mathcal{A})$ is the set of ground policies defined over states in s_ϕ, and $s \in s_\phi$ is shorthand for $s \in \{s' \mid \phi(s') = s_\phi, s' \in \mathcal{S}\}$. We denote $\mathcal{O}_\phi$ as any non-empty set that 1) contains only ϕ-relative options, and 2) contains at least one option that initiates in each $s_\phi \in \mathcal{S}_\phi$.

Intuitively, these options initiate in exactly one abstract state and terminate when the option policy leaves the abstract state. We henceforth denote $(\phi, \mathcal{O}_\phi)$ as a state abstraction paired with a set of ϕ-relative options.

Example. Consider the classical Four Rooms domain pictured in Figure 1a. Suppose further that the state abstraction ϕ turns each room into an abstract state. Then any ϕ-relative option in this domain would be one that initiates anywhere in one of the rooms and terminates as soon as the agent leaves that room. The only degree of flexibility in grounding a set of ϕ-relative options for the given ϕ, then, is which *policies* are associated with each option, and how many options are available in each abstract state. If, for instance, the optimal policy π^* were chosen for an option in the top right room, but the uniform random policy were available everywhere else, how might that impact the overall suboptimality of the policies induced by the abstraction? Our main result (Theorem 1) clarifies the precise conditions under which ε-optimal policies are representable under different abstractions.



(a) Assignment of options to each s_ϕ via $\pi_{\mathcal{O}_\phi}$.

$$\pi^{\Downarrow}_{\mathcal{O}_\phi}(s) = \begin{cases} \pi_{o_1}(s), & s \in \\ \pi_{o_2}(s), & s \in \\ \pi_{o_3}(s), & s \in \\ \pi_{o_4}(s), & s \in \end{cases}$$

(b) Construction of $\pi^{\Downarrow}_{\mathcal{O}_\phi}$.

Figure 1: Grounding policy $\pi_{\mathcal{O}_\phi}$ to $\pi^{\Downarrow}_{\mathcal{O}_\phi}$.

To analyze the value loss of these pairs, we first show that any $(\phi, \mathcal{O}_\phi)$ gives rise to an abstract policy over $\mathcal{S}_\phi$ and $\mathcal{O}_\phi$ that induces a unique policy in the original MDP (over the entire state space). Critically, this property does not hold for arbitrary options due to their semi-Markovian nature.

Proofs of all introduced Theorems and Remarks are presented in Appendix A.

Remark 1. *Every deterministic policy defined over abstract states and ϕ-relative options, $\pi_{\mathcal{O}_\phi} : \mathcal{S}_\phi \to \mathcal{O}_\phi$, induces a unique Markov policy in the ground MDP, $\pi^{\Downarrow}_{\mathcal{O}_\phi} : \mathcal{S} \to \Delta(\mathcal{A})$. We let $\Pi_{\mathcal{O}_\phi}$ denote the set of abstract policies representable by the pair $(\phi, \mathcal{O}_\phi)$, and $\Pi^{\Downarrow}_{\mathcal{O}_\phi}$ denote the corresponding set of policies in the original MDP.*

This remark gives us a means of translating a policy over ϕ-relative options into a policy over the original state and action space, $\mathcal{S}$ and $\mathcal{A}$. Consequently, we can define the value loss associated with a set of options paired with a state abstraction: every $(\phi, \mathcal{O}_\phi)$ pair yields a set of policies in the original MDP, $\Pi^{\Downarrow}_{\mathcal{O}_\phi}$. We define the value loss of $(\phi, \mathcal{O}_\phi)$ as the value loss of the best policy in this set.

Definition 5 ($(\phi, \mathcal{O}_\phi)$-Value Loss): *The value loss of $(\phi, \mathcal{O}_\phi)$ is the smallest degree of sub-optimality achievable:*

$$L(\phi, \mathcal{O}_\phi) := \min_{\pi_{\mathcal{O}_\phi} \in \Pi_{\mathcal{O}_\phi}} \left|\left| V^* - V^{\pi^{\Downarrow}_{\mathcal{O}_\phi}} \right|\right|_\infty . \tag{2}$$

Note that this notion of value loss is *not* well defined for options in general, since they induce a semi-MDP: there is no well-formed *ground* value function of a policy over options, but rather, a semi-Markov value function. As a simple illustration, consider a ground state s_g, two options o_1 and o_2 (either of which could be executing in s_g), and a policy $\pi_{\phi,o}$ over abstract states and options. It could be that o_1 or o_2 is currently executing when s_g is entered or that either option has just terminated, requiring $\pi_{\phi,o}$ to select a new option. Each of these three cases induces a distinct value $V^{\pi_{\phi,o}}(s_g)$ which is then difficult to distill into a single ground value function. This is a key reason to restrict attention to ϕ-relative options, each of which retains structure that couples with the corresponding state abstraction ϕ to yield value functions in the ground MDP.

3.1 Main Results

We next show how different classes of ϕ-relative options can represent near-optimal policies. We define an option class by a predicate $\lambda : \mathcal{O}_\phi \mapsto \{0, 1\}$, and say that a set of ϕ-relative options $\mathcal{O}_\phi$ belongs to the class $\mathcal{O}_{\phi,\lambda}$ if and only if $\lambda(\mathcal{O}_\phi) = 1$.

We begin by summarizing the four new ϕ-relative option classes, drawing inspiration from other forms of abstraction (Dean and Givan, 1997; Li et al., 2006; Jiang et al., 2015a; Abel et al., 2016; Ravindran and Barto, 2004; Nachum et al., 2019). For each class, we will refer to the *optimal* option in s_ϕ, $o^*_{s_\phi}$, as the ϕ-relative option which initiates in s_ϕ and executes π^* until termination. These classes were chosen as they closely parallel existing properties studied in the literature. The four classes are as follows:

1. *Similar Q^* Functions:* In each s_ϕ, there is at least one option o that has similar Q^* to $o^*_{s_\phi}$.
2. *Similar Models:* In each s_ϕ, there is at least one option o that has a similar multi-time model (Precup and Sutton, 1998) to $o^*_{s_\phi}$.
3. *Similar k-Step Distributions:* In each s_ϕ, there is at least one option o that has a similar k-step termination state distribution to $o^*_{s_\phi}$, based off the hierarchical construction introduced by Nachum et al. (2019). Loss bounds will only apply to goal-based MDPs.
4. *Approximate MDP Homomorphisms:* We show that any deterministic $\pi_{\mathcal{O}_\phi}$ can encode an MDP homomorphism. The MDP homomorphism option class is defined by a guarantee on the quality of the resulting homomorphism.

Our main result establishes the bounded value loss of pairs $(\phi, \mathcal{O}_\phi)$ where $\mathcal{O}_\phi$ belongs to any of these four classes, and the size of the bound depends on the degree of approximation (ε_Q; ε_R, ε_T; τ; and ε_r, ε_p).

Theorem 1. *(Main Result) For any ϕ, the four introduced classes of ϕ-relative options satisfy:*

$$L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \frac{\varepsilon_Q}{1-\gamma}, \tag{3}$$

$$L(\phi, \mathcal{O}_{\phi,M_\varepsilon}) \leq \frac{\varepsilon_R + |\mathcal{S}|\varepsilon_T \text{RMAX}}{(1-\gamma)^2}, \tag{4}$$

$$L(\phi, \mathcal{O}_{\phi,\tau}) \leq \frac{\tau\gamma|\mathcal{S}|}{(1-\gamma)^2}, \tag{5}$$

$$L(\phi, \mathcal{O}_{\phi,H}) \leq \frac{2}{1-\gamma}\left(\varepsilon_r + \frac{\gamma \text{RMAX}}{1-\gamma}\frac{\varepsilon_p}{2}\right), \tag{6}$$

where RMAX *is an upper bound on the reward function, the $L(\phi, \mathcal{O}_{\phi,\tau})$ bound holds in goal-based MDPs and the other three hold in any finite MDP.*

Observe that when the approximation parameters are zero, many of the bounds collapse to 0 as well. This illustrates the trade-off made between the amount of knowledge used to construct the abstractions and the degree of optimality ensured. Further note that the value loss of the state abstraction does not appear in any of the above bounds—indeed, ϕ will *implicitly* affect the value loss as a function of the diameter of each abstract state.

We now present each class in full technical detail. As stated, the first two classes guarantee ε closeness of values and models respectively. More concretely:

Similar Q^*-Functions ($\mathcal{O}_{\phi,Q^*_\varepsilon}$): The ε-similar Q^* predicate defines an option class where:

$$\lambda(\mathcal{O}_\phi) \equiv \tag{7}$$
$$\forall_{s_\phi \in \mathcal{S}_\phi} \exists_{o \in \mathcal{O}_\phi} : \max_{s \in s_\phi} |Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, o)| \leq \varepsilon_Q,$$

where

$$Q^*_{s_\phi}(s, o) := R(s, \pi_o(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s' \mid s, \pi_o(s)) \left(\mathbb{1}(s' \in s_\phi) Q^*_{s_\phi}(s', o) + \mathbb{1}(s' \notin s_\phi) V^*(s')\right). \tag{8}$$

This Q-function describes the expected return of starting in state s, executing a ϕ-relative option o until leaving $\phi(s)$, then following the optimal policy thereafter.

More generally, this class of $\phi, \mathcal{O}_\phi$ pairs captures all cases where each abstract state has at least one option that is useful. Note that the identity state abstraction paired with the degenerate set of options that exactly encodes the execution of each primitive action will necessarily be an instance of this class.

Similar Models ($\mathcal{O}_{\phi,M_\varepsilon}$): The ε-similar T and R predicate defines an option class where:

$$\lambda(\mathcal{O}_\phi) \equiv \forall_{s_\phi \in \mathcal{S}_\phi} \exists_{o \in \mathcal{O}_\phi} : \quad (9)$$
$$\left|\left| T^{s'}_{s,o^*_{s_\phi}} - T^{s'}_{s,o} \right|\right|_\infty \leq \varepsilon_T \text{ and } \left|\left| R_{s,o^*_{s_\phi}} - R_{s,o} \right|\right|_\infty \leq \varepsilon_R,$$

where $R_{s,o}$ and $T^{s'}_{s,o}$ are shorthand for the reward model and multi-time model of Sutton et al. (1999). Roughly, this class states that there is at least one option in each abstract state that behaves similarly to the optimal option in that abstract state, $o^*_{s_\phi}$, throughout its execution in the abstract state.

We next derive two classes of ϕ-relative options based on abstraction formalisms from existing literature. The first is based on the hierarchical construction introduced by Nachum et al. (2019), while the second shows that ϕ-relative options can describe an MDP homomorphism (Ravindran and Barto, 2004).

Similar k-Step Distributions ($\mathcal{O}_{\phi,\tau}$): Let $\mathbb{P}(s', k \mid s, o)$ denote the probability of option o terminating in s' after k steps, given that it initiated in s. We define this class by the following predicate:

$$\lambda(\mathcal{O}_\phi) \equiv \forall_{s_\phi \in \mathcal{S}_\phi} \exists_{o \in \mathcal{O}_\phi} \forall_k : \quad (10)$$
$$\max_{s \in s_\phi, s' \in \mathcal{S}} |\mathbb{P}(s', k \mid s, o^*_{s_\phi}) - \mathbb{P}(s', k \mid s, o)| \leq \tau.$$

For the next class definition we first define the one-step abstract transition and reward functions for a ϕ-relative option o:

$$T_\phi(s'_\phi \mid s_\phi, o) = \sum_{s \in s_\phi} w(s) \sum_{s' \in s'_\phi} T(s' \mid s, \pi_o(s)), \quad (11)$$

$$R_\phi(s_\phi, o) = \sum_{s \in s_\phi} w(s) R(s, \pi_o(s)), \quad (12)$$

where $w(s)$ is any valid weighting function with $\sum_{s \in s_\phi} w(s) = 1$. Next, we define the quantities of Ravindran and Barto (2004), related to the simulation lemma of Kearns and Singh (2002):

$$K_p = \max_{s \in \mathcal{S}, a \in \mathcal{A}} \sum_{s_\phi \in \mathcal{S}_\phi} \left| \sum_{s' \in s_\phi} T(s' \mid s, a) - T_\phi(s_\phi \mid \phi(s), \pi_{\mathcal{O}_\phi}(\phi(s))) \right|,$$

$$K_r = \max_{s \in \mathcal{S}, a \in \mathcal{A}} |R(s, a) - R_\phi(\phi(s), \pi_{\mathcal{O}_\phi}(\phi(s)))|. \quad (13)$$

These capture the maximum discrepancy between the model of the ground MDP and the model of the induced abstract MDP defined according to $(\phi, \mathcal{O}_\phi)$. Then, we define the class as follows.

Approximate MDP Homomorphisms ($\mathcal{O}_{\phi,H}$): This class requires that $\mathcal{O}_\phi$ represents policies that induce good approximate homomorphisms:

$$\lambda(\mathcal{O}_\phi) \equiv \forall\ \pi_{\mathcal{O}_\phi} \in \Pi_{\mathcal{O}_\phi} : K_p \leq \varepsilon_p \text{ and } K_r \leq \varepsilon_r.$$

These four classes constitute four sufficient conditions for $(\phi, \mathcal{O}_\phi)$ pairs to yield bounded value loss. It is useful, however, to identify not just sufficient conditions, but also necessary. To this end, we next establish one necessary condition of all (globally) value preserving $(\phi, \mathcal{O}_\phi)$ classes.

Theorem 2. *For any $(\phi, \mathcal{O}_\phi)$ pair with $L(\phi, \mathcal{O}_\phi) \leq \eta$, there exists at least one option per abstract state that is η-optimal in Q-value. Precisely, if $L(\phi, \mathcal{O}_\phi) \leq \eta$, then:*

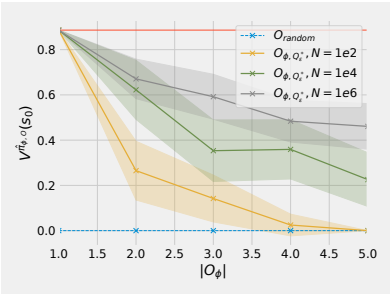
$$\forall_{s_\phi \in \mathcal{S}_\phi} \forall_{s \in s_\phi} \exists_{o \in \mathcal{O}_\phi} : Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, o) \leq \eta. \quad (14)$$

This theorem tells us that for any agent acting using our joint abstraction framework, if there exists an abstract state for which there is not an η-optimal option, then the agent cannot represent a globally near-optimal policy.

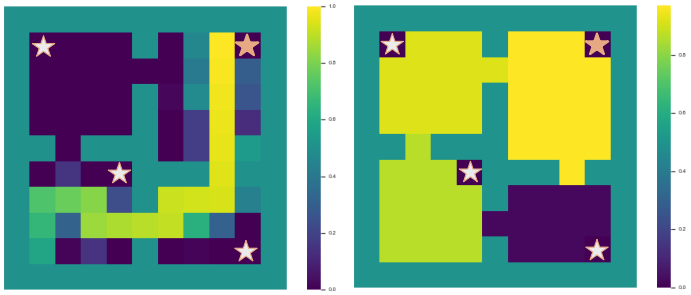
3.2 Experiment

We next conduct a simple experiment to corroborate the findings of our analysis, and to test whether value preserving options enable simple RL algorithms to find near-optimal policies. The experiment illustrates an important property of one of the introduced $\phi, \mathcal{O}_\phi$ classes, and is organized as follows. We first construct a $\phi, \mathcal{O}_\phi$ pair belonging to the $\mathcal{O}_{\phi,Q^*_\varepsilon}$ class using dynamic programming. We give this pair to an instance of Double Q-Learning (Hasselt, 2010) with a varied sample budget N. The environment is a Four Rooms grid world MDP, with a single goal location in the top right and start location in the bottom left. The state abstraction ϕ maps each state into one of four abstract states, denoting each of the four rooms. We vary both the number of options added per abstract state ($|\mathcal{O}_\phi|$) and the sample budget of Double Q (N), and present the value of the policy discovered by the final episode for each setting of $|\mathcal{O}_\phi|$ and N.

Results are presented in Figure 2a. First, note that with only one option per abstract state, Double Q can trivially find a near-optimal policy, even with a small sample budget. This aligns with our theory: the included options preserve value, and so any assignment



(a) Learning With Value Preserving Options



(b) Learned Value Functions for Q-Learning (left) and Q-Learning with $\phi, \mathcal{O}_\phi$ (right)

Figure 2: Empirical evidence that the $\phi, \mathcal{O}_\phi$ pairs from Theorem 1 preserve value (left) and comparison of the learned value function with Q-Learning run in the ground MDP and run in the abstract MDP of $\phi, \mathcal{O}_\phi$ (right). On the left, the y-axis denotes the value of the policy learned by Double Q-Learning for the given sample budget (N) and option set, averaged over 25 runs with 95% confidence interval. Optimal behavior is shown in red. On the right, the stars indicate potential goal locations (those that were used for constructing the options), with the red star the currently active one. Brighter color indicates a higher estimated value, with purple a value of 0.

of options to abstract states will yield a near-optimal policy. In contrast, if randomly chosen options are used instead (shown in blue, labeled as O_{random}), the learning algorithm fails to find a good policy even with a high sample budget ($N = 1e6$ was used). Second, we find that as the number of options increases, the added branching factor causes Double Q to find a lower value policy with the same number of samples. However, by Theorem 1 we know each set of options preserves value; as the sample budget increases we see that the value of the discovered policy tends toward optimal. In short, the $\phi, \mathcal{O}_\phi$ pairs defined by Theorem 1 do in fact preserve value, but will also affect the sample budget required to find a good policy. We foresee the combination of value preserving abstractions with those that lower learning complexity (see recent work by Brunskill and Li (2014); Fruit et al. (2017)) as a key direction for future work.

We further visualize the learned value function of Q-Learning with and without ϕ-relative options after the same sample budget, depicted in Figure 2b. Notably, since ϕ-relative options update entire blocks of states, we see large regions of the state space with the same learned value function. Conversely, Q-Learning only tends to explore (and estimate the values of) a narrow region of the state space. The visual highlights this important qualitative difference between learning with and without abstractions. More experiments, details, and a link to the code are found in Appendix B.

4 HIERARCHICAL ABSTRACTION

We next present an extension of Theorem 1 that applies to hierarchies consisting of $(\phi, \mathcal{O}_\phi)$ pairs. We show the value loss compounds linearly if we construct a hierarchy using algorithms that generate a well-behaved ϕ and $\mathcal{O}_\phi$. To do so, we require two definitions and additional notation (see Table 1 in Appendix A for a summary of hierarchy notation). We first define a *hierarchy* as n sets of $(\phi, \mathcal{O}_\phi)$ pairs.

Definition 6 ($(\phi, \mathcal{O}_\phi)$-Hierarchy): *A depth n hierarchy, denoted H_n, is a list of n state abstractions, $\phi^{(n)}$, and a list of n sets of ϕ-relative options, $\mathcal{O}_\phi^{(n)}$. The components $(\mathcal{I}, \beta, \pi)$ of each of the i-th set of options, $\mathcal{O}_{\phi,i}$ are defined over the $(i-1)$-th abstract state space $\mathcal{S}_{\phi,i-1} = \{\phi_{i-1}(\phi_{i-2}(\dots\phi_1(s)\dots)) \mid s \in \mathcal{S}\}$.*

We next introduce additional notation to refer to values, states, options, and policies at each level of the hierarchy. We denote $\pi_n : \mathcal{S}_{\phi,n} \to \mathcal{O}_{\phi,n}$ as the level n policy encoded by the hierarchy, with Π_n the space of all policies encoded in this way. We let $s_i := \phi^i(s) = \phi_i(\dots\phi_1(s)\dots)$, with s a state in the ground MDP. We further denote V_i as the i-th level's value function, defined as follows for some ground state s:

$$V_i^\pi(s) := V_i^\pi\left(\phi^i(s)\right) = V_i^\pi\left(s_i\right) = \max_{o \in \mathcal{O}_i} \left(R_i\left(s_i, o\right) + \sum_{s' \in \mathcal{S}_i} T_i\left(s' \mid s_i, o\right) V_i^\pi(s') \right), \quad (15)$$

where:

$$R_i(s_i, o) := \sum_{s_{i-1} \in s_i} w_i(s_{i-1}) R_{s_{i-1}, o} \quad (16)$$

$$T_i(s_i' \mid s_i, o) := \sum_{s_{i-1} \in s_i} \sum_{s_{i-1}' \in \mathcal{S}_{i-1}} w_i(s_{i-1}) T_{s_{i-1}, o}^{s_{i-1}'} \quad (17)$$

where again $R_{s,o}$ and $T_{s,o}^{s'}$ are defined according to the multi-time model (Sutton et al., 1999), $s_i \in \mathcal{S}_{\phi,i}$ is a

level i state resulting from $\phi^i(s)$, and w_i is an aggregation weighting function for level i. Note that V_0 is the ground value function, which we refer to as V for simplicity.

4.1 Hierarchy Analysis

We now extend Theorem 1 to hierarchies of arbitrary depth, building on two key observations. First, any policy π_n represented at the top level of a hierarchy H_n also has a unique Markov policy in the ground MDP, which we denote $\pi_n^{\Downarrow}$ (in contrast to $\pi_n^{\downarrow}$, which moves the level n policy to level $n-1$). We summarize this fact in the following remark:

Remark 2. *Every deterministic policy π_i defined by the i-th level of a hierarchy, H_n, induces a unique policy in the ground MDP, which we denote $\pi_i^{\Downarrow}$.*

To be precise, note that $\pi_i^{\downarrow}$ specifies the level i policy π_i mapped into level π_{i-1}, whereas $\pi_i^{\Downarrow}$ refers to the policy at π_i mapped into π_0. The second key insight is that we can extend our same notion of value loss from $(\phi, \mathcal{O}_\phi)$ pairs to hierarchies, H_n.

Definition 7 (H_n-Value Loss): *The value loss of a depth n hierarchy H_n is the smallest degree of suboptimality across all policies representable at the top level of the hierarchy:*

$$L(H_n) := \min_{\pi_n \in \Pi_n} \left|\left| V^* - V^{\pi_n^{\Downarrow}} \right|\right|_\infty . \tag{18}$$

This quantity denotes how suboptimal the best hierarchical policy is in the ground MDP. Therefore, the guarantee we present expresses a condition on *global* optimality rather than *recursive* or *hierarchical* optimality (Dietterich, 2000a).

We next show that there exist value-preserving hierarchies by bounding the above quantity for well constructed hierarchies. To prove this result, we require two assumptions.

Assumption 1. *The value function is consistent throughout the hierarchy. That is, for every level of the hierarchy $i \in [1:n]$, for any policy π_i over states $\mathcal{S}_{\phi,i}$ and options $\mathcal{O}_{\phi,i}$, there is a small $\kappa \in \mathbb{R}_{\geq 0}$ such that:*

$$\max_{s \in \mathcal{S}} \left| V_{i-1}^{\pi_i^{\downarrow}}\left(\phi^{i-1}(s)\right) - V_i^{\pi_i}\left(\phi^i(s)\right) \right| \leq \kappa \tag{19}$$

Assumption 2. *Subsequent levels of the hierarchy can represent policies similar in value to the best policy at the previous level. That is, for every $i \in [1:n-1]$, letting $\pi_i^\diamond = \arg\min_{\pi_i \in \Pi_i} ||V_0^* - V_0^{\pi_i^{\Downarrow}}||_\infty$, there is a small $\ell \in \mathbb{R}_{\geq 0}$ such that:*

$$\min_{\pi_{i+1}^{\downarrow} \in \Pi_{i+1}^{\downarrow}} \left|\left| V_i^{\pi_i^\diamond} - V_i^{\pi_{i+1}^{\downarrow}} \right|\right|_\infty \leq \ell. \tag{20}$$

We strongly suspect that both assumptions are true given the right choice of state abstractions, options, and methods of constructing abstract MDPs. These two assumptions (along with Theorem 1) give rise to hierarchies that can represent near-optimal behavior. We present this fact through the following theorem:

Theorem 3. *Consider two algorithms: 1) A_ϕ: given an MDP M, outputs a ϕ, and 2) $A_{\mathcal{O}_\phi}$: given M and a ϕ, outputs a set of options $\mathcal{O}$ such that there are constants κ and ℓ for which Assumption 1 and Assumption 2 are satisfied.*

Then, by repeated application of A_ϕ and $A_{\mathcal{O}_\phi}$, we can construct a hierarchy of depth n such that

$$L(H_n) \leq n(\kappa + \ell). \tag{21}$$

This theorem offers a clear path for extending the guarantees of ϕ-relative options beyond the typical two-timescale setup observed in recent work (Bacon et al., 2017; Nachum et al., 2019) to fully realize the benefits of (multi-level) hierarchical abstraction (Levy et al., 2019). Moreover, both Assumption 1 and Assumption 2 are sufficient—together with ϕ-relative options that satisfy Theorem 1—to construct a hierarchy with low value loss. We conclude that algorithms for leveraging hierarchies may want to explicitly search for structures that satisfy our assumptions: 1) value function smoothness up and down the hierarchy, and 2) policy richness at each level of the hierarchy.

5 DISCUSSION

We proved which state-action abstractions are guaranteed to preserve representation of high value policies. To do so, we introduced ϕ-relative options, a simple but expressive formalism for combining state abstractions with options. Under this formalism, we proposed four classes of ϕ-relative options with bounded value loss. Lastly, we proved that under mild conditions, pairs of state-action abstractions can be recursively combined to induce hierarchies that also possess near-optimality guarantees.

We take these results to serve as a concrete path toward principled abstraction discovery and use in RL. To realize this goal, abstractions need to both preserve good behavior, and lower sample complexity. Our results thus far focus on the representation of near-optimal policies, which is a key condition for ensuring agents can eventually behave well. However, learning agents also need to represent functions of intermediate quality as they learn. Thus, an important direction for future work is to clarify which kinds of abstractions ensure that near-optimal policies are easily learnable.

Acknowledgements

We would like to thank Cam Allen, Philip Amortila, Will Dabney, Mark Ho, George Konidaris, Ofir Nachum, Scott Niekum, Silviu Pitis, and Balaraman Ravindran for insightful discussions, and the anonymous reviewers for their thoughtful feedback. This work was supported by a grant from the DARPA L2M program and an ONR MURI grant.

References

David Abel, D. Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the International Conference on Machine Learning*, 2016.

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael L. Littman. State abstractions for lifelong reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2018.

David Abel, Dilip Arumugam, Kavosh Asadi, Yuu Jinnai, Michael L. Littman, and Lawson L.S. Wong. State abstraction as compression in apprenticeship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Ankit Anand, Aditya Grover, Parag Singla, et al. A novel abstraction framework for online planning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2015.

David Andre and Stuart Russell. State abstraction for programmable reinforcement learning agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2002.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Aijun Bai and Stuart Russell. Efficient reinforcement learning with hierarchies of machines by leveraging internal transitions. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017.

Andrew G. Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1-2):41–77, 2003.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684, 1957.

Emma Brunskill and Lihong Li. PAC-inspired option discovery in lifelong reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2014.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Pablo Samuel Castro and Doina Precup. Automatic construction of temporally extended actions for MDPs using bisimulation metrics. In *Proceedings of the European Workshop on Reinforcement Learning*, 2011.

Kamil Ciosek and David Silver. Value iteration with options and state aggregation. *ICAPS Workshop on Planning and Learning*, 2015.

Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, 1993.

Thomas Dean and Robert Givan. Model minimization in Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1997.

Richard Dearden and Craig Boutilier. Abstraction and approximate decision-theoretic planning. *Artificial Intelligence*, 89(1):219–283, 1997.

Thomas G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 2000a.

Thomas G. Dietterich. State abstraction in MAXQ hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, 2000b.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2004.

Norman Ferns, Pablo Samuel Castro, Doina Precup, and Prakash Panangaden. Methods for computing state similarity in Markov decision processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2006.

Bennett L Fox. Discretizing dynamic programs. *Journal of Optimization Theory and Applications*, 11(3): 228–234, 1973.

Ronan Fruit and Alessandro Lazaric. Exploration–exploitation in MDPs with options. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2017.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. Regret minimization in MDPs with options without prior knowledge. In *Advances in Neural Information Processing Systems*, 2017.

Robert Givan, Sonia Leach, and Thomas Dean. Bounded parameter Markov decision processes. In *European Conference on Planning*. Springer, 1997.

Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option:

Learning options with a deliberation cost. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Anna Harutyunyan, Will Dabney, Diana Borsa, Nicolas Heess, Remi Munos, and Doina Precup. The termination critic. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

Hado Van Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, 2010.

Jesse Hostetler, Alan Fern, and Thomas G. Dietterich. State aggregation in MCTS. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014.

Marcus Hutter. Extreme state aggregation beyond MDPs. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 2014.

Nan Jiang, Satinder Singh, and Richard Lewis. Improving uct planning via approximate homomorphisms. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, 2014.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2015a.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2015b.

Yuu Jinnai, Jee Won Park, David Abel, and George Konidaris. Discovering options for exploration by minimizing cover time. In *Proceedings of the International Conference on Machine Learning*, 2019.

Nicholas K. Jong and Peter Stone. State abstraction discovery from irrelevant state variables. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005.

Nicholas K. Jong and Peter Stone. Hierarchical model-based reinforcement learning: R- MAX+MAXQ. In *Proceedings of the International Conference on Machine Learning*, 2008.

Nicholas K. Jong, Todd Hester, and Peter Stone. The utility of temporal abstraction in reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2008.

Anders Jonsson and Andrew G. Barto. Automated state abstraction for options using the U-tree algorithm. In *Advances in Neural Information Processing Systems*, 2001.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

George Konidaris. Constructing abstraction hierarchies using a skill-symbol loop. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2016.

George Konidaris and Andrew G. Barto. Building portable options: Skill transfer in reinforcement learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2007.

George Konidaris and Andrew G. Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. In *Advances in Neural Information Processing Systems*, 2009.

George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research*, 2018.

Kim G. Larsen and Arne Skou. Bisimulation through probabilistic testing. *Information and computation*, 94(1):1–28, 1991.

Lucas Lehnert, Romain Laroche, and Harm van Seijen. On value function representation of long horizon problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. In *Proceedings of the International Conference on Learning Representations*, 2019.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*, 2006.

Marios C Machado, Marc G Bellemare, and Michael Bowling. A laplacian framework for option discovery in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *International Conference on Machine Learning*, pages 543–551, 2013.

Sultan Javed Majeed and Marcus Hutter. Performance guarantees for homomorphisms beyond Markov de-

cision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Timothy A. Mann and Shie Mannor. Scaling up approximate value iteration with options: Better policies with fewer iterations. In *Proceedings of the International Conference on Machine Learning*, 2014.

Timothy A. Mann, Shie Mannor, and Doina Precup. Approximate value iteration with temporally extended actions. *Journal of Artificial Intelligence Research*, 2015.

Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Near-optimal representation learning for hierarchical reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2019.

Maillard Odalric-Ambrym, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2013.

Ronald Ortner, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit, and Odalric-Ambrym Maillard. Regret bounds for learning state representations in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

Ronald Parr and Stuart Russell. Reinforcement learning with hierarchies of machines. In *Advances in Neural Information Processing Systems*, 1998.

Doina Precup and Richard S. Sutton. Multi-time models for temporally abstract planning. In *Advances in Neural Information Processing Systems*, 1998.

Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Balaraman Ravindran. *SMDP homomorphisms: An algebraic approach to abstraction in semi Markov decision processes*. PhD thesis, University of Massachusetts Amherst, 2003.

Balaraman Ravindran and Andrew G. Barto. Model minimization in hierarchical reinforcement learning. In *International Symposium on Abstraction, Reformulation, and Approximation*, pages 196–211. Springer, 2002.

Balaraman Ravindran and Andrew G. Barto. Relativized options: Choosing the right transformation. In *Proceedings of the International Conference on Machine Learning*, 2003a.

Balaraman Ravindran and Andrew G. Barto. SMDP homomorphisms: An algebraic approach to abstraction in semi-Markov decision processes. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2003b.

Balaraman Ravindran and Andrew G. Barto. Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes. In *Proceedings of the International Conference on Knowledge Based Computer Systems*, 2004.

Özgür Şimşek and Andrew G. Barto. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2004.

Özgür Şimşek and Andrew G. Barto. Skill characterization based on betweenness. In *Advances in Neural Information Processing Systems*, 2009.

Satinder Singh, Tommi Jaakkola, and Michael I. Jordan. Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, 1995.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 1999.

Jonathan Taylor, Doina Precup, and Prakash Panagaden. Bounding performance loss in approximate MDP homomorphisms. In *Advances in Neural Information Processing Systems*, 2008.

Saket Tiwari and Philip S. Thomas. Natural option critic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Nicholay Topin, Nicholas Haltmeyer, Shawn Squire, John Winder, James MacGlashan, et al. Portable option discovery for automated learning transfer in object-oriented Markov decision processes. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2015.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1-3):59–94, 1996.

Benjamin Van Roy. Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2):234–244, 2006.

Ward Whitt. Approximations of dynamic programs, i. *Mathematics of Operations Research*, 3(3):231–243, 1978.

Ward Whitt. Approximations of dynamic programs, ii. *Mathematics of Operations Research*, 4(2):179–185, 1979.

Marco Wiering and Jürgen Schmidhuber. HQ-learning. *Adaptive Behavior*, 6(2):219–246, 1997.

Value Preserving State-Action Abstractions (Appendix)

David Abel
Brown University

Nathan Umbanhowar
Brown University

Khimya Khetarpal
Mila-McGill University

Dilip Arumugam
Stanford University

Doina Precup
Mila-McGill University

Michael L. Littman
Brown University

We here present proofs of each introduced theoretical result (Appendix A) along with additional experiments and implementation details (Appendix B). All of our code is publicly available for extension and reproduction.[1]

A Proofs

In this section we provide proofs of each introduced result.

Remark 1. *Every deterministic policy defined over abstract states and ϕ-relative options, $\pi_{\mathcal{O}_\phi} : \mathcal{S}_\phi \to \mathcal{O}_\phi$, induces a unique Markov policy in the ground MDP, $\pi^{\Downarrow}_{\mathcal{O}_\phi} : \mathcal{S} \to \Delta(\mathcal{A})$. We denote by $\Pi_{\mathcal{O}_\phi}$ the set of abstract policies representable by the pair $(\phi, \mathcal{O}_\phi)$, and let $\Pi^{\Downarrow}_{\mathcal{O}_\phi}$ be the corresponding set of policies in the original MDP.*

Proof. Consider an arbitrary deterministic policy $\pi_{\mathcal{O}_\phi}$. By definition, this policy assigns one option to each abstract state. Let $\mathcal{O}_\pi$ denote the set of options this policy assigns.

By construction of ϕ-relative options, for every ground state $s \in \mathcal{S}$ there is one unique option $o_{\phi(s)} \in \mathcal{O}_\pi$ that can be executed in s.

Therefore, we construct a policy $\pi^{\Downarrow}_{\mathcal{O}_\phi}$ as the combination of option policies in $\mathcal{O}_\pi$. Specifically, letting $\pi_{o_{\phi(s)}}$ denote the option policy of the option in $\mathcal{O}_\pi$ that is assigned to $\phi(s)$:

$$\pi^{\Downarrow}_{\mathcal{O}_\phi}(s) = \pi_{o_{\phi(s)}}(s) \tag{22}$$

□

.....

Theorem 1. *(Main Result) For any ϕ, the four introduced classes of ϕ-relative options satisfy:*

$$L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \frac{\varepsilon_Q}{1-\gamma}, \qquad L(\phi, \mathcal{O}_{\phi,M_\varepsilon}) \leq \frac{\varepsilon_R + |\mathcal{S}|\varepsilon_T \text{RMax}}{(1-\gamma)^2}, \tag{23}$$

$$L(\phi, \mathcal{O}_{\phi,\tau}) \leq \frac{\tau\gamma|\mathcal{S}|}{(1-\gamma)^2}, \qquad L(\phi, \mathcal{O}_{\phi,H}) \leq \frac{2}{1-\gamma}\left(\varepsilon_r + \frac{\gamma \text{RMax}}{1-\gamma}\frac{\varepsilon_p}{2}\right), \tag{24}$$

where the $L(\phi, \mathcal{O}_{\phi,\tau})$ bound holds in goal-based MDPs and the other three hold in any MDP.

We prove this claim using four separate proofs, each targeting one class.

[1] https://github.com/david-abel/vpsa_aistats2020

Proof. $(L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \frac{\varepsilon_Q}{1-\gamma})$

Consider $L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) = \min_{\pi^{\Downarrow}_{\mathcal{O}_\phi} \in \Pi^{\Downarrow}_{\mathcal{O}_\phi}} \max_{s \in \mathcal{S}} |V^*(s) - V^{\pi^{\Downarrow}_{\mathcal{O}_\phi}}(s)|$. Since $V^*(s) \geq V^\pi(s)$ for all π, we henceforth drop the absolute value for convenience.

To proceed, we recall that $o^*_{s_\phi}$ is the ϕ-relative option that executes π^* in every state and terminates when it leaves the abstract state s_ϕ:

$$o^*_{s_\phi} := \forall_{s \in \mathcal{S}} : \langle \mathcal{I}(s) \equiv \phi(s) = s_\phi, \tag{25}$$

$$\beta(s) \equiv \phi(s) \neq s_\phi, \tag{26}$$

$$\pi(s) = \pi^*(s) \rangle. \tag{27}$$

Note that since $o^*_{s_\phi}$ always chooses actions according to π^*, that $Q^*_{s_\phi}(s, o^*_{s_\phi}) = V^*(s)$ (where $Q^*_{s_\phi}$ is defined according to Equation 8).

Then, by the Q^*_ε predicate, we can construct a policy over abstract states and options $\mu_{\mathcal{O}_\phi} \in \Pi_{\mathcal{O}_\phi}$ with the following property:

$$\forall_{s_\phi \in \mathcal{S}_\phi, s \in s_\phi} : Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \varepsilon_Q. \tag{28}$$

Note that $\mu_{\mathcal{O}_\phi}(s_\phi)$ outputs an option. As in Equation 28, we henceforth denote $s_\phi = \phi(s)$ and correspondingly $s'_\phi = \phi(s')$.

Then it must be the case that

$$L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \max_{s \in \mathcal{S}} V^*(s) - V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s). \tag{29}$$

Let $Q^*_t(s, o)$ denote the expected discounted reward of executing option o, then executing t options under $\mu_{\mathcal{O}_\phi}$, then following the optimal policy thereafter. Note that

$$\lim_{t \to \infty} Q^*_t(s, \mu_{\mathcal{O}_\phi}(s_\phi)) = V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s), \tag{30}$$

because $Q^*_t(s, \mu_{\mathcal{O}_\phi}(s_\phi))$ is the expected discounted reward of executing $t+1$ options under $\mu_{\mathcal{O}_\phi}$, then following the optimal policy thereafter.

We next show by induction on t that

$$\max_{s \in \mathcal{S}} V^*(s) - V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s) = \max_{s \in \mathcal{S}} \lim_{t \to \infty} V^*(s) - Q^*_t(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \frac{\varepsilon_Q}{1-\gamma}. \tag{31}$$

In particular, we wish to show that

$$\forall_{t \in \mathbb{N}} : \max_{s \in \mathcal{S}} V^*(s) - Q^*_t(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{t} \varepsilon_Q \gamma^i. \tag{32}$$

(Base Case)
When $t = 0$, for all $s \in \mathcal{S}$,

$$Q^*_0(s, \mu_{\mathcal{O}_\phi}(s_\phi)) = Q^*_{s_\phi}(s, \mu_{\mathcal{O}_\phi}(s_\phi)), \tag{33}$$

because both quantities represent the expected discounted reward of executing the option $\mu_{\mathcal{O}_\phi}(s_\phi)$ then following the optimal policy thereafter. It follows that

$$\max_{s \in \mathcal{S}} V^*(s) - Q^*_0(s, \mu_{\mathcal{O}_\phi}(s_\phi)) = \max_{s \in \mathcal{S}} V^*(s) - Q^*_{s_\phi}(s, \mu_{\mathcal{O}_\phi}(s_\phi)), \tag{34}$$

$$= \max_{s \in \mathcal{S}} Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, \mu_{\mathcal{O}_\phi}(s_\phi)), \tag{35}$$

$$\leq \varepsilon_Q, \tag{36}$$

$$= \sum_{i=0}^{0} \varepsilon_Q \gamma^0, \tag{37}$$

where the inequality holds by definition of $\mu_{\mathcal{O}_\phi}$.

(Inductive Case)
We assume as the inductive hypothesis that

$$\max_{s\in\mathcal{S}} V^*(s) - Q^*_k(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{k} \varepsilon_Q \gamma^i, \tag{38}$$

and want to show that

$$\max_{s\in\mathcal{S}} V^*(s) - Q^*_{k+1}(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i. \tag{39}$$

To begin, fix $s \in \mathcal{S}$ and consider

$$V^*(s) - Q^*_{k+1}(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \tag{40}$$

$$= V^*(s) - \left(R_o(s, \mu_{\mathcal{O}_\phi}(s_\phi)) + \sum_{s'\in\mathcal{S}} T_o(s'|s, \mu_{\mathcal{O}_\phi}(s_\phi)) Q^*_k(s', \mu_{\mathcal{O}_\phi}(s'_\phi)) \right) \tag{41}$$

$$= V^*(s) - R_o(s, \mu_{\mathcal{O}_\phi}(s_\phi)) - \sum_{s'\in\mathcal{S}} T_o(s'|s, \mu_{\mathcal{O}_\phi}(s_\phi)) Q^*_k(s', \mu_{\mathcal{O}_\phi}(s'_\phi)) \tag{42}$$

where R_o and T_o indicate the reward and multi-time option models from Sutton et al. (1999).

Now, subtract and add $\sum_{s'\in\mathcal{S}} T_o(s'|s, \mu_{\mathcal{O}_\phi}(s_\phi)) V^*(s')$:

$$= \quad V^*(s) - R_o(s, \mu_{\mathcal{O}_\phi}(s_\phi)) - \sum_{s'\in\mathcal{S}} T_o(s' \mid s, \mu_{\mathcal{O}_\phi}(s_\phi)) V^*(s') \tag{43}$$

$$+ \sum_{s'\in\mathcal{S}} T_o(s' \mid s, \mu_{\mathcal{O}_\phi}(s_\phi)) V^*(s') - \sum_{s'\in\mathcal{S}} T_o(s' \mid s, \mu_{\mathcal{O}_\phi}(s_\phi)) Q^*_k(s', \mu_{\mathcal{O}_\phi}(s'_\phi))$$

$$= \quad V^*(s) - Q^*_{s_\phi}(s, \mu_{\mathcal{O}_\phi}(s_\phi)) + \sum_{s'\in\mathcal{S}} T_o(s' \mid s, \mu_{\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\mathcal{O}_\phi}(s'_\phi)) \right] \tag{44}$$

$$= \quad Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, \mu_{\mathcal{O}_\phi}(s_\phi)) + \sum_{s'\in\mathcal{S}} T_o(s' \mid s, \mu_{\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\mathcal{O}_\phi}(s'_\phi)) \right] \tag{45}$$

$$\leq \quad \varepsilon_Q + \sum_{s'\in\mathcal{S}} T_o(s' \mid s, \mu_{\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\mathcal{O}_\phi}(s'_\phi)) \right], \tag{46}$$

by definition of $\mu_{\mathcal{O}_\phi}$. Continuing, we have that:

$$= \quad \varepsilon_Q + \sum_{s'\in\mathcal{S}} \sum_{n=1}^{\infty} \mathbb{P}(s', n|s, \mu_{\mathcal{O}_\phi}(s_\phi)) \gamma^n \left[V^*(s') - Q^*_k(s', \mu_{\mathcal{O}_\phi}(s'_\phi)) \right] \tag{47}$$

$$\leq \quad \varepsilon_Q + \sum_{s'\in\mathcal{S}} \sum_{n=1}^{\infty} \mathbb{P}(s', n|s, \mu_{\mathcal{O}_\phi}(s_\phi)) \gamma^n \sum_{i=0}^{k} \varepsilon_Q \gamma^i, \tag{48}$$

by the inductive hypothesis. Then:

$$= \quad \varepsilon_Q + \gamma \sum_{s'\in\mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}(s', n+1|s, \mu_{\mathcal{O}_\phi}(s_\phi)) \gamma^n \sum_{i=0}^{k} \varepsilon_Q \gamma^i \tag{49}$$

$$= \quad \varepsilon_Q + \gamma \sum_{i=0}^{k} \varepsilon_Q \gamma^i \sum_{s'\in\mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}(s', n+1|s, \mu_{\mathcal{O}_\phi}(s_\phi)) \gamma^n \tag{50}$$

$$\leq \quad \varepsilon_Q + \gamma \sum_{i=0}^{k} \varepsilon_Q \gamma^i \cdot 1 \tag{51}$$

$$= \quad \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i, \tag{52}$$

since $\mathbb{P}(s', n+1|s, \mu_{\mathcal{O}_\phi}(s_\phi))$ is a probability distribution and γ is less than 1.

All together, we've shown that $V^*(s) - Q^*_{k+1}(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i$ for all $s \in \mathcal{S}$, which implies that

$$\max_{s\in\mathcal{S}} V^*(s) - Q^*_{k+1}(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i, \tag{53}$$

as desired.

It follows by induction that

$$\forall_{t\in\mathbb{N}} : \max_{s\in\mathcal{S}} V^*(s) - Q^*_t(s, \mu_{\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{t} \varepsilon_Q \gamma^i. \tag{54}$$

Therefore,

$$\begin{aligned} L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) &\leq \max_{s\in\mathcal{S}} V^*(s) - V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s) && (55)\\ &= \max_{s\in\mathcal{S}} \lim_{t\to\infty} V^*(s) - Q^*_t(s, \mu_{\mathcal{O}_\phi}(s_\phi)) && (56)\\ &\leq \lim_{t\to\infty} \sum_{i=0}^{t} \varepsilon_Q \gamma^i && (57)\\ &= \frac{\varepsilon_Q}{1-\gamma}, && (58) \end{aligned}$$

which completes the proof.

□

.....

Proof. $(L(\phi, \mathcal{O}_{\phi,M_\varepsilon}) \leq \frac{\varepsilon_R + |\mathcal{S}|\varepsilon_T \textsc{VMax}}{1-\gamma})$

We show that this class is a subclass of the $\mathcal{O}_{\phi,Q^*_\varepsilon}$ class. Therefore, it stands to show that, given our class definition, there exists an option in every abstract state that is near-optimal in Q-value.

Fix $s \in \mathcal{S}$. Let $s_\phi = \phi(s)$. By the M_ε predicate, there exists an option $o \in \mathcal{O}_\phi$ such that

$$||T^{s'}_{s,o^*_{s_\phi}} - T^{s'}_{s,o}||_\infty \leq \varepsilon_T \text{ and } ||R_{s,o^*_{s_\phi}} - R_{s,o}||_\infty \leq \varepsilon_R. \tag{59}$$

Now, we consider the difference in optimal Q-values between $o^*_{s_\phi}$ and o. We first have that:

$$\begin{aligned} Q^*_{s_\phi}(s, o) &= R(s, \pi_o(s)) + \gamma \sum_{s'\in\mathcal{S}} T(s' \mid s, \pi_o(s)) \left(\mathbb{1}(s' \in s_\phi) Q^*_{s_\phi}(s', o) + \mathbb{1}(s' \notin s_\phi) V^*(s') \right) \\ &= R_o(s, o) + \sum_{s'\in\mathcal{S}} T_o(s'|s, o) V^*(s'), \end{aligned} \tag{60}$$

with R_o and T_o denoting the reward model and multi-time model of Sutton et al. (1999).
By symmetry,

$$Q^*_{s_\phi}(s, o^*_{s_\phi}) = R_o(s, o^*_{s_\phi}) + \sum_{s'\in\mathcal{S}} T_o(s'|s, o^*_{s_\phi}) V^*(s'). \tag{61}$$

Therefore,

$$\begin{aligned}|Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, o)| &= |R_o(s, o^*_{s_\phi}) - R_o(s, o) + \sum_{s' \in \mathcal{S}} T_o(s'|s, o^*_{s_\phi}) V^*(s') - \\ &\qquad \sum_{s' \in \mathcal{S}} T_o(s'|s, o) V^*(s')| \\ &\leq |R_o(s, o^*_{s_\phi}) - R_o(s, o)| + |\sum_{s' \in \mathcal{S}} \Big(T_o(s'|s, o^*_{s_\phi}) - T_o(s'|s, o) \Big) V^*(s')| \\ &\leq |R_o(s, o^*_{s_\phi}) - R_o(s, o)| + \sum_{s' \in \mathcal{S}} |T_o(s'|s, o^*_{s_\phi}) - T_o(s'|s, o)||V^*(s')| \\ &\leq \varepsilon_R + |\mathcal{S}| \varepsilon_T \text{VMAX},\end{aligned} \tag{62}$$

by the model similarity assumption. We have now shown that any option with near-optimal models has a near-optimal Q-value with $\varepsilon_Q = \varepsilon_R + |\mathcal{S}| \varepsilon_T \text{VMAX}$. Therefore, by the previous result,

$$L(\phi, O_{\phi, M_\varepsilon}) \leq \frac{\varepsilon_R + |\mathcal{S}| \varepsilon_T \text{VMAX}}{1 - \gamma}. \tag{63}$$

□

.....

Proof. $(L(\phi, \mathcal{O}_{\phi,\tau}) \leq \frac{\tau\gamma|\mathcal{S}|}{(1-\gamma)^2})$

We first state rigorously our definition of a goal-based MDP.

Definition 8 (Goal-based MDP): *A goal-based MDP is an MDP with some number of goal states, denoted $\mathcal{S}_G \subseteq \mathcal{S}$. The reward function is such that $R(s, a) = 1$ if $s \in \mathcal{S}_G$, $R(s, a) = 0$ otherwise, and the episode terminates after receiving a reward in a goal state. Furthermore, we assume that each goal state exists in its own abstract state: $s \neq s_G \Rightarrow \phi(s_G) \neq \phi(s)$, where $s_G \in \mathcal{S}_G, s \in \mathcal{S}$.*

We show that this class is a subclass of the $\mathcal{O}_{\phi, Q^*_\varepsilon}$ class in goal-based MDPs. In particular, it stands to show that given our class definition, there exists an option in every abstract state that is near-optimal in Q-value.

First, note that in the abstract states containing a goal state, any option is optimal since $R(s, a) = 1$ regardless of action. Therefore, we restrict our attention to an arbitrary $s \in \mathcal{S} \setminus \mathcal{S}_G$, fixing $s_\phi = \phi(s)$. Let o be an option available in s_ϕ such that $\max_{s \in s_\phi, s' \in \mathcal{S}} |\mathbb{P}(s', k \mid s, o^*_{s_\phi}) - \mathbb{P}(s', k \mid s, o)| \leq \tau$, by the option class definition. Then

$$Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, o) \tag{64}$$

$$= R_o(s, o^*_{s_\phi}) + \sum_{s' \in \mathcal{S}} T_o(s'|s, o^*_{s_\phi}) V^*(s') - R_o(s, o) - \sum_{s' \in \mathcal{S}} T_o(s'|s, o) V^*(s') \tag{65}$$

$$= \sum_{s' \in \mathcal{S}} \Big[T_o(s'|s, o^*_{s_\phi}) - T_o(s'|s, o) \Big] V^*(s'), \tag{66}$$

where we drop the R_o terms since $s \notin \mathcal{S}_G$, each goal state has its own abstract state, and $R(s, a) = 0$ for $s \notin \mathcal{S}_G$. Continuing, we have that

$$Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, o) = \sum_{s' \in \mathcal{S}} \left[\sum_{k=1}^{\infty} |\mathbb{P}(s', k \mid s, o^*_{s_\phi}) - \mathbb{P}(s', k \mid s, o)| \gamma^k \right] V^*(s), \tag{67}$$

writing out the multi-time model. This implies that

$$Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, o) \leq \sum_{s' \in \mathcal{S}} \frac{\tau\gamma}{1 - \gamma} V^*(s). \tag{68}$$

Now, note that $V^*(s') = \sum_{s_G \in \mathcal{S}_G} \sum_{t=0}^{\infty} p(s_G, t \mid s', \pi^*)\gamma^t$ in a goal-based MDP, where $p(s_G, t \mid s', \pi^*)$ is the probability of being in state s_G after t timesteps, starting from s' and following π^*. Indeed, this gives that $V^*(s') \leq 1$ since $p(s_G, t \mid s', \pi^*)$ is a probability distribution and γ is less than one. Therefore,

$$Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, o) \leq \frac{\tau\gamma|\mathcal{S}|}{1-\gamma}. \tag{69}$$

We have shown that there exists an option, o, in any abstract state that is near-optimal in Q-value, with $\varepsilon_Q = \frac{\tau\gamma|\mathcal{S}|}{1-\gamma}$. Therefore, by the $\mathcal{O}_{\phi, Q^*_\varepsilon}$ bound,

$$L(\phi, \mathcal{O}_{\phi,\tau}) \leq \frac{\tau\gamma|\mathcal{S}|}{(1-\gamma)^2}, \tag{70}$$

as desired.

□

.....

Proof. $\left(L(\phi, \mathcal{O}_{\phi,H}) \leq \frac{2}{1-\gamma}\left(\varepsilon_r + \frac{\gamma \text{RMax}}{1-\gamma}\frac{\varepsilon_p}{2}\right)\right)$

We prove this result by illustrating the connection between our formalisms and the work of Ravindran and Barto (2004). To do so, we first restate their definition of an approximate homomorphism.

Definition 9 (Approximate Homomorphism (Ravindran and Barto (2004))): *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 $a \in A_s$:*

$$P'\left(f(s), g_s(a), f\left(s'\right)\right) = \sum_{(q,b)\in[(s,a)]_h} w_{qb} \sum_{s''\in[s']_f} P\left(q, b, s''\right) \tag{71}$$

$$R'\left(f(s), g_s(a)\right) = \sum_{(q,b)\in[(s,a)]_h} w_{qb}\ R(q,b), \tag{72}$$

where $[(s,a)]_h$ denotes the preimage of $h((s,a))$, $[s']_f$ denotes the preimage of $f(s')$, and $\sum_{(q,b)\in[(s,a)]_h} w_{qb} = 1$. Furthermore, Ψ and Ψ' denote the sets of admissible state-action pairs in the ground and abstract MDP respectively. Based on Ψ and Ψ', A_s denotes the set of actions available in state s of the ground MDP, and $A'_{f(s)}$ denotes the set of abstract actions available in state $f(s)$ of the abstract MDP.

We now illustrate how our definitions of ϕ, R_ϕ, T_ϕ with respect to a given $\pi_{\mathcal{O}_\phi}$ induce an approximate homomorphism. First, note that our ϕ precisely corresponds to their definition of f, a state abstraction. Then, fix $s_\phi \in \mathcal{S}_\phi$, and let $A'_{s_\phi} = \{\pi_{\mathcal{O}_\phi}(s_\phi)\}$ with $g_s(a) = \pi_{\mathcal{O}_\phi}(s_\phi) \forall_{s \in s_\phi}\ \forall_{a \in A}$.

We now consider our definitions of T_ϕ and R_ϕ:

$$T_\phi(s'_\phi \mid s_\phi, o) = \sum_{s\in s_\phi} w(s) \sum_{s'\in s'_\phi} T(s'|s, \pi_o(s)) \qquad R_\phi(s_\phi, o) = \sum_{s\in s_\phi} w(s) R(s, \pi_o(s)), \tag{73}$$

We note that these are precisely an instance of P' and R' as defined above, with $w_{qb} = 0$ whenever $b \neq \pi_o(q)$. We write $w(s)$ to denote this choice of weighting function, which depends only on the action prescribed by π_o. We select this choice of weighting function (as opposed to a weighting dependent on all available actions) in order to faithfully represent the 1-step behavior of executing an option in the abstract MDP.

By these connections, a deterministic policy $\pi_{\mathcal{O}_\phi}$ over ϕ-relative options coupled with our choice of weighting function defines an approximate homomorphism. We further adapt their definitions of K_p and K_r to our notational setting, which describe the maximum discrepancy in models between the ground and abstract MDPs.

$$K_p = \max_{s\in\mathcal{S},a\in\mathcal{A}} \sum_{s_\phi\in\mathcal{S}_\phi} \Big| \sum_{s'\in s_\phi} T(s'|s,a) - T_\phi(s_\phi|\phi(s),\pi_{\mathcal{O}_\phi}(\phi(s)))\Big|, \tag{74}$$

$$K_r = \max_{s\in\mathcal{S},a\in\mathcal{A}} |R(s,a) - R_\phi(\phi(s),\pi_{\mathcal{O}_\phi}(\phi(s)))|. \tag{75}$$

The main theorem of Ravindran and Barto (2004) guarantees that the value loss of the optimal policy in the abstract MDP $\mathcal{M}'$ is upper-bounded by

$$\frac{2}{1-\gamma}\left(K_r + \frac{\gamma}{1-\gamma}\delta_{r'}\frac{K_p}{2}\right),$$

where $\delta_{r'}$ is upper-bounded by RMax. Let $\mu_{\mathcal{O}_\phi} \in \Pi_{\mathcal{O}_\phi}$ denote the optimal policy in the abstract MDP. By our option class definition, all abstract policies $\pi_{\mathcal{O}_\phi} \in \Pi_{\mathcal{O}_\phi}$ induce homomorphisms with bounded K_p, K_r, so, in particular, $\mu_{\mathcal{O}_\phi}$ has bounded K_p, K_r. Then:

$$L(\phi,\mathcal{O}_{\phi,H}) = \min_{\pi_{\mathcal{O}_\phi}\in\Pi_{\mathcal{O}_\phi}} \left|\left|V^* - V^{\pi^{\Downarrow}_{\mathcal{O}_\phi}}\right|\right|_\infty \tag{76}$$

$$\leq \left|\left|V^* - V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}\right|\right|_\infty \tag{77}$$

$$\leq \frac{2}{1-\gamma}\left(K_r + \frac{\gamma}{1-\gamma}\delta_{r'}\frac{K_p}{2}\right) \tag{78}$$

$$\leq \frac{2}{1-\gamma}\left(\varepsilon_r + \frac{\gamma \text{RMax}}{1-\gamma}\frac{\varepsilon_p}{2}\right), \tag{79}$$

as desired. □

.....

Theorem 2. *For any $(\phi,\mathcal{O}_\phi)$ pair with $L(\phi,\mathcal{O}_\phi) \leq \eta$, there must exist at least one option per abstract state that is η-optimal in Q-value. Precisely, if $L(\phi,\mathcal{O}_\phi) \leq \eta$, then:*

$$\forall_{s_\phi\in\mathcal{S}_\phi}\forall_{s\in s_\phi}\exists_{o\in\mathcal{O}_\phi} : Q^*_{s_\phi}(s,o^*) - Q^*_{s_\phi}(s,o) \leq \eta. \tag{80}$$

Proof. Let $\mu_{\mathcal{O}_\phi} = \arg\min_{\pi_{\mathcal{O}_\phi}\in\Pi_{\mathcal{O}_\phi}} \left|\left|V^* - V^{\pi^{\Downarrow}_{\mathcal{O}_\phi}}\right|\right|_\infty$.
Suppose, for contradiction, that there exists an abstract state s_ϕ for which there is no η-optimal option in $\mathcal{O}_\phi$. Then it must be the case that

$$Q^*_{s_\phi}(s,o^*) - Q^*_{s_\phi}(s,\mu_{\mathcal{O}_\phi}(s)) > \eta \tag{81}$$

for some $s \in s_\phi$.

By, $Q^*_{s_\phi}(s,o^*) = V^*(s)$, this implies that

$$V^*(s) - Q^*_{s_\phi}(s,\mu_{\mathcal{O}_\phi}(s)) > \eta. \tag{82}$$

Then, note that $Q^*_{s_\phi}(s,\mu_{\mathcal{O}_\phi}(s)) \geq V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s)$ because $Q^*_{s_\phi}$ describes the expected return of executing option $\mu_{\mathcal{O}_\phi}(s)$, then switching to optimal behavior, whereas $V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}$ describes the expected return of executing $\mu_{\mathcal{O}_\phi}(s)$ then continuing to execute options according to $\mu_{\mathcal{O}_\phi}$.
Noticing that $V^*(s) \geq Q^*_{s_\phi}(s,\mu_{\mathcal{O}_\phi}(s)) \geq V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s)$, we have that

$$V^*(s) - V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s) > \eta. \tag{83}$$

This implies that

$$L(\phi, \mathcal{O}_\phi) = \min_{\pi_{\mathcal{O}_\phi} \in \Pi_{\mathcal{O}_\phi}} \left|\left| V^* - V^{\pi^{\Downarrow}_{\mathcal{O}_\phi}} \right|\right|_\infty \tag{84}$$

$$= V^*(s) - V^{\mu^{\Downarrow}_{\mathcal{O}_\phi}}(s) \tag{85}$$

$$> \eta, \tag{86}$$

which contradicts the premise. Therefore, it must be true that

$$\forall_{s_\phi \in \mathcal{S}_\phi} \forall_{s \in s_\phi} \exists_{o \in \mathcal{O}_\phi} : Q^*_{s_\phi}(s, o^*) - Q^*_{s_\phi}(s, o) \leq \eta. \tag{87}$$

□

.....

ϕ	A state abstraction function.
$\mathcal{O}_\phi$	A set of ϕ-relative options.
$L(\phi, \mathcal{O}_\phi)$	The value loss of the $\phi, \mathcal{O}_\phi$ pair.
$\pi_{\mathcal{O}_\phi}$	A policy that maps each abstract state to an option.
$\pi^{\Downarrow}_{\mathcal{O}_\phi}$	A policy over $\mathcal{S}$ and $\mathcal{A}$, induced by $\pi_{\mathcal{O}_\phi}$.
H_n	A hierarchy of depth n, denoting the pair of lists $(\phi^{(n)}, \mathcal{O}^{(n)}_\phi)$.
$\phi^{(n)}$	A list of n state abstractions, where $\phi_i : \mathcal{S}_{\phi,i-1} \to \mathcal{S}_{\phi,i}$.
ϕ_i	The i-th state abstraction in a list $\phi^{(n)}$.
ϕ^i	The result of applying the first i state abstractions to s, $\phi_i(\dots \phi_1(s) \dots)$.
$\mathcal{S}_{\phi,i}$	The i-th abstract state space.
s_i	A state belonging to $\mathcal{S}_{\phi,i}$
V^π_i	Value of level i under policy π, defined according to R_i and T_i.
$\mathcal{O}_{\phi,i}$	Options at level i, with each component defined over states in $\mathcal{S}_{\phi,i-1}$.
R_i	The reward function of level i.
T_i	The reward function of level i.
π_i	The policy over level i of the hierarchy such that $\pi_i : \mathcal{S}_i \to \mathcal{O}_{\phi,i}$.
$\pi^{\downarrow}_i$	A policy over $\mathcal{S}_{\phi,i-1}$ and $\mathcal{O}_{\phi,i-1}$, induced by π_i.
$\pi^{\Downarrow}_i$	A policy over $\mathcal{S}$ and $\mathcal{A}$, induced by π_i.

Table 1: Abstraction notation.

A.1 Hierarchical Analysis

Our aim is to generalize Theorem 1 arbitrary hierarchies, H_n. To do so, we make two key observations. First, any policy π_n represented at the top level of a hierarchy H_n also has a unique Markov policy in the ground MDP, which we denote $\pi^{\Downarrow}_n$ (in contrast to $\pi^{\downarrow}_n$, which moves the level n policy to level $n-1$). We summarize this fact in the following lemma:

Remark 2. *Every deterministic policy π_i defined according to the i-th level of a hierarchy, H_n, induces a unique policy in the ground MDP, which we denote $\pi^{\Downarrow}_i$.*

To be precise, note that $\pi^{\downarrow}_i$ specifies the level i policy π_i mapped into level π_{i-1}, whereas $\pi^{\Downarrow}_i$ refers to the policy at π_i mapped into π_0. For further details regarding notion, see Table 1.

The second key insight is that the same notion of value loss from $\phi, \mathcal{O}_\phi$ pairs can be extended to hierarchies, H_n.

Definition 10 (H_n-Value Loss): *The value loss of a depth n hierarchy H_n is the smallest degree of suboptimality across all policies representable at the top level of the hierarchy:*

$$L(H_n) := \min_{\pi_n \in \Pi_n} \left|\left| V^* - V^{\pi_n^\Downarrow} \right|\right|_\infty. \tag{88}$$

Note that the above value functions are the value function in the original MDP; this bound evaluates how suboptimal the best hierarchical policy is *in the ground MDP*. We next show that there exist value-preserving hierarchies by bounding the above quantity for well constructed hierarchies. To prove this result, we require two assumptions.

Assumption 1. *The value function is consistent throughout the hierarchy. That is, for every level of the hierarchy $i \in [1:n]$, for any policy π_i over states $\mathcal{S}_{\phi,i}$ and options $\mathcal{O}_{\phi,i}$, there is a small $\kappa \in \mathbb{R}_{\geq 0}$ such that:*

$$\max_{s \in \mathcal{S}} \left| V_{i-1}^{\pi_i^\downarrow}\left(\phi^{i-1}(s)\right) - V_i^{\pi_i}\left(\phi^i(s)\right) \right| \leq \kappa \tag{89}$$

Assumption 2. *Subsequent levels of the hierarchy can represent policies similar in value to the best policy at the previous level. That is, for every $i \in [1:n-1]$, letting $\pi_i^\diamond = \arg\min_{\pi_i \in \Pi_i} ||V_0^* - V_0^{\pi_i^\Downarrow}||_\infty$, there is a small $\ell \in \mathbb{R}_{\geq 0}$ such that:*

$$\min_{\pi_{i+1}^\downarrow \in \Pi_{i+1}^\downarrow} \left|\left| V_i^{\pi_i^\diamond} - V_i^{\pi_{i+1}^\downarrow} \right|\right|_\infty \leq \ell. \tag{90}$$

We strongly suspect that both assumptions are true given the right choice of state abstractions, options, and methods of constructing abstract MDPs. As some motivating evidence, a claim closely related to Assumption 1 is proven by Abel et al. (2016) as Claim 1, and Assumption 2 is of similar structure to our own Theorem 1. Regardless, these two assumptions (along with Theorem 1) give rise to hierarchies that can represent near-optimal behavior. We present this fact through the following theorem:

Theorem 3. *Consider two algorithms: 1) A_ϕ: given an MDP M, outputs a ϕ, and 2) $A_{\mathcal{O}_\phi}$: given M and a ϕ, outputs a set of options $\mathcal{O}$ such that there are constants κ and ℓ for which Assumption 1 and Assumption 2 are satisfied. Then, by repeated application of A_ϕ and $A_{\mathcal{O}_\phi}$, we can construct a hierarchy of depth n such that*

$$L(H_n) \leq n(\kappa + \ell). \tag{91}$$

Proof. We present the proof of the bound for a two level hierarchy, but the same strategy generalizes to n levels via induction.

Let ℓ be the known upper bound for $L(\phi, \mathcal{O})$. Then:

By Theorem 1:	$\min_{\pi_1 \in \Pi_1} ||V_0^* - V_0^{\pi_1^\downarrow}||_\infty \leq \ell$
By Assumption 1:	$\forall_{\pi_1 \in \Pi_1} : ||V_0^{\pi_1^\downarrow} - V_1^{\pi_1}||_\infty \leq \kappa$
Letting $\pi_1^\diamond = \arg\min_{\pi_1 \in \Pi_1} ||V_0^* - V_0^{\pi_1^\downarrow}||_\infty$, by Assumption 2:	$\min_{\pi_2^\downarrow \in \Pi_2^\downarrow} ||V_1^{\pi_1^\diamond} - V_1^{\pi_2^\downarrow}||_\infty \leq \ell$
By Assumption 1	$\forall_{\pi_2^\downarrow \in \Pi_2^\downarrow} : ||V_1^{\pi_2^\downarrow} - V_0^{\pi_2^\Downarrow}||_\infty \leq \kappa$

Therefore, by the triangle inequality:

$$\min_{\pi_2 \in \Pi_2} ||V_0^* - V_0^{\pi_2^\Downarrow}||_\infty \leq 2\kappa + 2\ell. \tag{92}$$

□

In short: the right hierarchies, constructed out of $\phi, \mathcal{O}_\phi$ pairs, can also preserve value.

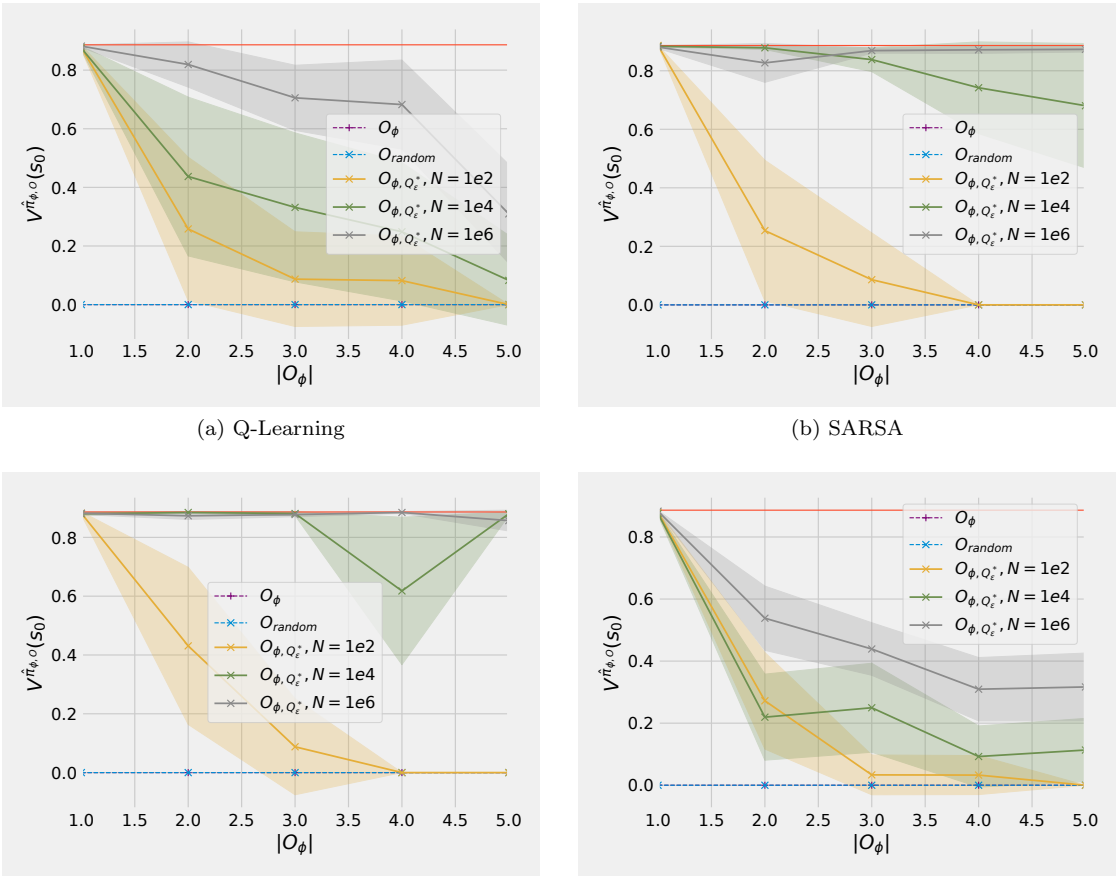
.....

B Experimental Details

We next provide further detail about the experiment described in Section 3.2.

The environment used is the Four Rooms grid world domain from Sutton et al. (1999). We place the start state in the bottom left corner and the goal state in the top right corner, with no slip probability. We experiment with Double Q-Learning (Hasselt, 2010), given access to different $\phi, \mathcal{O}_\phi$ pairs from the $\mathcal{O}_{\phi,Q^*_\varepsilon}$. We define the size of the option set as follows: the first option included for each abstract state is guaranteed to have at worst an ε-sub-optimal policy within the cluster, as defined in the proof of the bounded value loss for the class. To construct this policy, we explicitly create an ε suboptimal version of π^* via Lemma 2 of Arumugam et al. (2018). When $|\mathcal{O}| > 1$, we add options that execute the uniform random policy within the cluster, until it exits (and hence, terminates the option). Thus, the learning problem requires that the agent discovers options of each abstract state which belong to the near-optimal policy and learns to ignore others. We set the exploration parameter ϵ for Double Q to be 0.1, the learning rate α to be .05, and $\gamma = 0.95$, with no tuning.

We ran further variations of the experiment with other canonical RL algorithms, including Q-Learning (Watkins and Dayan, 1992), R-Max (Brafman and Tennenholtz, 2002), SARSA (Rummery and Niranjan, 1994), and Delayed Q-Learning (Strehl et al., 2006). Results are presented in Figure 3. Again, we find the same trend uncovered in the experiment with Double Q-Learning:



(a) Q-Learning

(b) SARSA

(c) Delayed Q-Learning

(d) R-Max

Figure 3: Learning with value preserving $\phi, \mathcal{O}_\phi$ pairs for different algorithms.

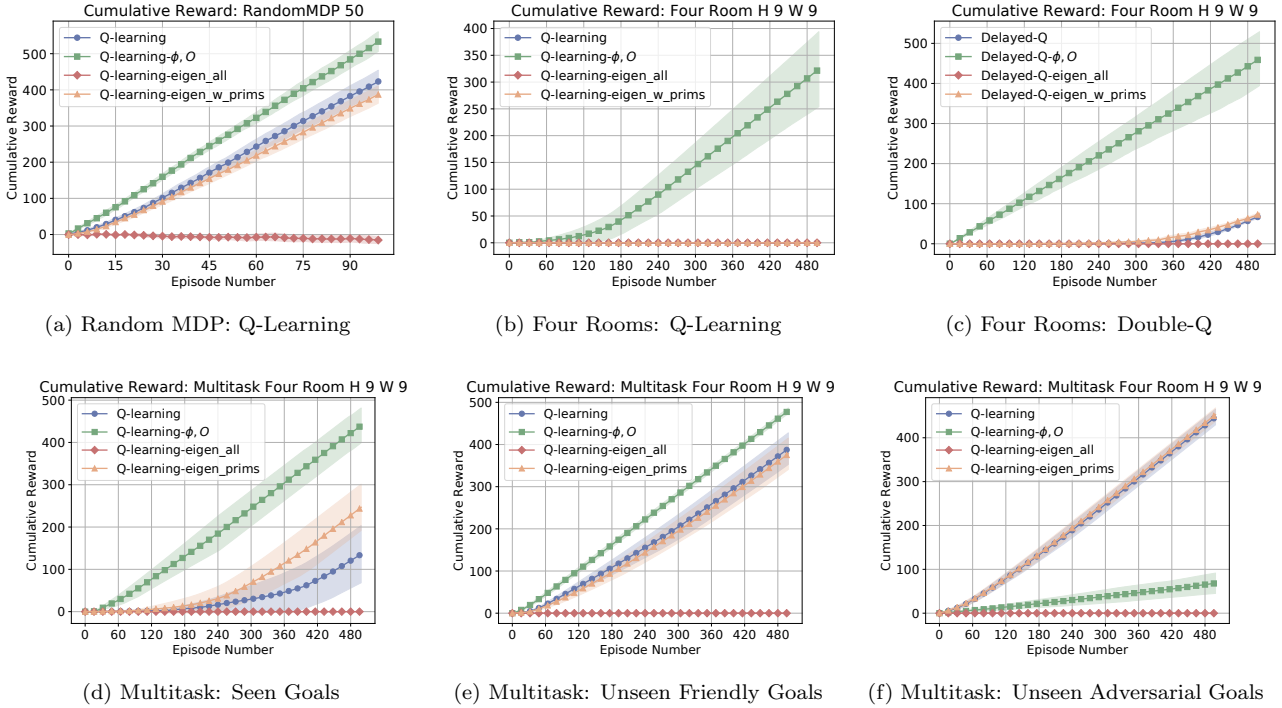
B.1 Experiment: Learning with $\phi, \mathcal{O}_\phi$

We next conduct learning experiments of two kinds: 1) single-task, and 2) multi-task, in both cases contrasting the sample efficiency of learning algorithms with and without value-preserving abstractions. We first construct pairs $(\phi, \mathcal{O}_\phi)$ prescribed by the $\phi, \mathcal{O}_{Q^*_\varepsilon}$ class with $\varepsilon = 0.05$. We conduct experiments in the classic Four Rooms MDP (Sutton et al., 1999) and in a random graph MDP. We test with two different algorithms: 1) Q-Learning (Watkins and Dayan, 1992), and 2) Delayed Q-Learning (Strehl et al., 2006). We ran each of the algorithms with and without pairs $(\phi, \mathcal{O}_\phi)$ from the option classes analyzed in Theorem 1.

We compare performance to learning algorithms on their own and given *eigenoptions*, which are chosen due to their capacity for effective exploration. As in the previous experiment, we test with two variants of eigenoptions: 1) -eigen_all, in which the primitive actions are removed and the options initiate in all states, and 2) -eigen_prims, in which the options are added to the primitive actions.

Single Task In the single-task experiments, we let each algorithm-abstraction pair interact with the Four Rooms MDP for 500 episodes with each episode consisting of 75 steps, and the Random MDP for 500 episodes with 25 steps per episode. We present the average cumulative reward achieved per episode across 10 runs with 95% confidence intervals.

Results for the Four Rooms experiments are presented in Figure 4b and Figure 4c. Unsurprisingly, we find that both learning algorithms are more sample efficient with value-preserving pairs $(\phi, \mathcal{O}_\phi)$, requiring a few episodes to learn a near-optimal policy (see Q-learning-ϕ, O and Delayed-Q-ϕ, O, both in green). In contrast, the baseline learning algorithms are unable to learn a reasonable policy even after around 400 episodes. The eigenoption variant shown in red further exposes the difficulty of value preservation: since the algorithm can only reason with the options, it is *never* able to find a good policy. Notably, the orange approach that includes primitive actions is able to also unable to learn, since it has to search through the fully policy space representable by the primitive actions. Results for the



(a) Random MDP: Q-Learning

(b) Four Rooms: Q-Learning

(c) Four Rooms: Double-Q

(d) Multitask: Seen Goals

(e) Multitask: Unseen Friendly Goals

(f) Multitask: Unseen Adversarial Goals

Figure 4: Results for single-task learning experiments in Four Rooms and the Random MDP (top) and multi-task learning experiments in Four Rooms (bottom).

Random MDP experiment are presented in Figure 4a; again, we find the value preserving abstractions are capable of supporting efficient learning of a high value policy. Here, eigenoptions paired with primitives achieves roughly the same learning speed as the baseline algorithm, while the variant without primitives can never learn a high-value policy.

Multitask In the multitask setting, we fixed a uniform random probability distribution over goal states. Then, we defined a state abstraction that picks out each goal state as its own abstract state, and otherwise groups each of the four rooms into four different abstract states, which we give to Q-Learning-(ϕ, O). At time step zero we sample a goal from the goal distribution and let each algorithm interact with the sampled MDP for 200 episodes, with each episode consisting of 50 steps. At the end of the 200 episodes, we reset each agent to *tabula rasa*, sample a new goal, and repeat. We present the mean cumulative reward achieved averaged over 50 samples from the distribution, with 95% confidence intervals. These results indicate how much the prior knowledge encoded by the abstractions improves sample efficiency over the entire distribution of goals. We again compare to the two variants of eigenoptions discussed in the single task experiment.

Results are presented in the bottom row of Figure Figure 4. We consider three distributions of goals: 1) *seen goals*, in which the agent constructs ϕ-relative options for goals it sees during learning (Figure 4d); 2) *unseen but friendly goals*, containing some goals the agent did not see during the construction of the options, but are close to those seen during the option construction (Figure 4e), and 3) *unseen but adversarial goals*, where the agent is faced with some goals not seen during construction of the options that are distant from those seen (Figure 4f). As expected, when the agent faces familiar goals, the abstraction-equipped learner is far more sample efficient than other approaches. Indeed, in under fifty episodes, Q-Learning-ϕ, O tends to find a high-value policy as seen in Figure 4d. Conversely, as the goals shift to being out of distribution, the improvement is less significant, as showcased by the drop in the green line's performance in Figure 4e. In the adversarial case, we construct goals that avoid representation by the ϕ-relative options the agent has constructed, thereby ensuring worse overall performance than the baseline learner. We find that the eigenoptions, given primitives, can learn faster than the baseline in some cases, and is typically competitive. Without primitives, however, eigenoptions can never discover a good policy, since no high-value policies can be represented.

B.2 Experiment: Value Loss

We next establish further empirical support of our main result by contrasting the value loss of basic abstraction types in small MDPs.

	Four Rooms	Lava Maze	Random	Hanoi	Taxi
$\max_{\pi \in \Pi_M} V^\pi(s_0)$	0.86	0.71	76.12	0.74	0.94
$\max_{\pi \in \Pi_{\mathcal{O}_\phi}} V^\pi(s_0)$	0.85	0.70	72.12	0.66	0.94

Table 2

In Table 2 we illustrate the value loss of our first class of value preserving ϕ-relative options ($\mathcal{O}_{\phi, Q^*_\varepsilon}$) in simple MDPs. Each row indicates the value (shown in blue) of the best policy representable using the policy space induced by the abstractions. As expected, the value preserving options can still represent a near-optimal policy in each MDP. For instance, in Four Rooms, $\Pi_{\mathcal{O}_\phi}$ achieves value of 0.85 compared to the optimal value of 0.86.

In short, we find support for our main theorem: value preserving ϕ-relative options can in fact preserve representation of near-optimal policies when reasoning only in terms of options. In most MDPs tested, the optimal policy over options only deviates from V^* by a small amount, as expected.

References

David Abel, D. Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the International Conference on Machine Learning*, 2016.

Dilip Arumugam, David Abel, Kavosh Asadi, Nakul Gopalan, Christopher Grimm, Jun Ki Lee, Lucas Lehnert, and Michael L. Littman. Mitigating planner overfitting in model-based reinforcement learning. *arXiv preprint arXiv:1812.01129*, 2018.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX—a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Hado Van Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, 2010.

Balaraman Ravindran and Andrew G. Barto. Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes. In *Proceedings of the International Conference on Knowledge Based Computer Systems*, 2004.

Gavin A. Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2006.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 1999.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.