VALUE PRESERVING STATE-ACTION ABSTRACTIONS

David Abel[1], **Nathan Umbanhowar**[1], **Khimya Khetarpal**[2], **Dilip Arumugam**[3], **Doina Precup**[2], **Michael L. Littman**[1]
{david_abel,umbanhowar}@brown.edu, khimya.khetarpal@mail.mcgill.ca, dilip@cs.stanford.edu, dprecup@cs.mcgill.ca, mlittman@cs.brown.edu
[1]Brown University, USA
[2]McGill University, CA
[3]Stanford University, USA

ABSTRACT

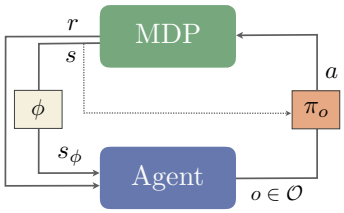
We here introduce combinations of state abstractions and options that preserve representation of near-optimal policies. We define ϕ-relative options, a general formalism for analyzing the value loss of options paired with a state abstraction, and prove that there exist classes of ϕ-relative options that preserve near-optimal behavior in any MDP. We conclude by proving that ϕ-relative options naturally induce a hierarchy, and that this hierarchy also preserves near-optimal behavior with value loss increasing as a function of the hierarchy's depth.

1 INTRODUCTION

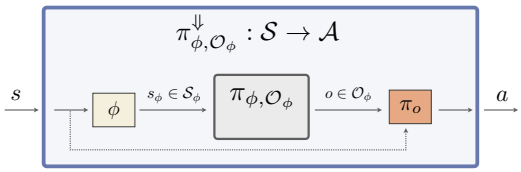
We here explore the role of *state* and *action* abstractions in the context of Reinforcement Learning (RL), as pictured in Figure 1a. Our objective is to clarify which combinations of state and action abstractions support near-optimal behavior in Markov Decision Processes (MDPs).

A state abstraction defines an aggregation function that translates the environmental state space $\mathcal{S}$ into $\mathcal{S}_\phi$, where usually $|\mathcal{S}_\phi| \ll |\mathcal{S}|$. With a smaller state space, learning algorithms can learn with less computation, space, and even samples (Dearden & Boutilier, 1997; Dietterich, 2000; Ravindran, 2003; Jong & Stone, 2005; Odalric-Ambrym et al., 2013; Hostetler et al., 2014; Jiang et al., 2015). However, throwing away information about the state space might destroy representation of good policies. An important direction for research is to clarify which state abstractions can preserve near-optimal behavior (Dean & Givan, 1997; Andre & Russell, 2002; Li et al., 2006; Hutter, 2014; Jiang et al., 2015; Abel et al., 2016; 2019).

We take an action abstraction to be a replacement of the actions of an MDP, $\mathcal{A}$, with a set of options (Sutton et al., 1999), $\mathcal{O}$, which encode long-horizon sequences of actions. Options are known to aid in transfer (Konidaris & Barto, 2007; Brunskill & Li, 2014; Topin et al., 2015), encourage better exploration (Bacon et al., 2017; Fruit & Lazaric, 2017; Machado et al., 2018; Tiwari & Thomas, 2019), and make planning more efficient (Mann & Mannor, 2014; Mann et al., 2015).



(a) State-Action Abstraction in RL.



(b) Forming a ground policy from a policy over abstract states and ϕ-relative options.

Figure 1: Reinforcement Learning with state abstraction and options: (a) an augmentation of the traditional RL loop wherein an agent reasons in terms of abstract states and chooses among options, and (b) the process for inducing $\pi^{\Downarrow}_{\phi,\mathcal{O}_\phi}$, a policy in the ground MDP, from a $(\phi, \mathcal{O}_\phi, \pi_{\phi,\mathcal{O}_\phi})$ triple.

The primary contribution of this work introduces combinations of state abstractions and options that preserve representation of near-optimal behavior. We define ϕ-relative options, a general formalism for analyzing the value loss of pairs $(\phi, \mathcal{O})$, and prove there are classes of ϕ-relative options that preserve near-optimal behavior in any MDP. We conclude by proving that this recursively yields a hierarchy that preserves near-optimal behavior under assumptions on the hierarchy's construction.

1.1 BACKGROUND

We first provide brief background on state abstractions and options.

Definition 1 (State Abstraction): *A state abstraction $\phi : \mathcal{S} \rightarrow \mathcal{S}_\phi$ maps each ground state, $s \in \mathcal{S}$ into an abstract state, $s_\phi \in \mathcal{S}_\phi$. We denote policies over abstract states as π_ϕ, defined as a mapping $\mathcal{S}_\phi \rightarrow \mathcal{A}$.*

Critically, a policy over abstract states induces a unique policy over ground states:

Remark 1. *Any deterministic policy defined over abstract states, $\pi_\phi : \mathcal{S}_\phi \rightarrow \mathcal{A}$ induces a unique policy in the original MDP. We denote this policy as $\pi_\phi^{\Downarrow}$, and the space of all policies representable in this manner as $\Pi_\phi^{\Downarrow}$.*

For each $s \in \mathcal{S}$, we may pass it through the abstraction to yield $s_\phi = \phi(s)$. To specify an action, we then query $\pi_\phi(s_\phi)$. Using this mapping process we can evaluate a given abstract policy, π_ϕ, by the value of its induced ground policy, $\pi_\phi^{\Downarrow}$. We now define the sub-optimality induced by a given state abstraction ϕ.

Definition 2 (ϕ-Value Loss): *The value loss associated with a state abstraction ϕ denotes the degree of sub-optimality attained by applying the best abstract policy. More formally:*

$$L(\phi) := \min_{\pi_\phi \in \Pi_\phi} \left|\left| V^* - V^{\pi_\phi^{\Downarrow}} \right|\right|_\infty \tag{1}$$

Next we introduce options, a popular formalism for augmenting the action space of an agent.

Definition 3 (Option (Sutton et al., 1999)): *An option $o \in \mathcal{O}$ is a triple $\langle \mathcal{I}_o, \beta_o, \pi_o \rangle$, 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 $\pi_o : \mathcal{S} \rightarrow \mathcal{A}$ is a deterministic policy prescribed by the option o.*

Options define abstract actions; the three components indicate where the option o can be executed ($\mathcal{I}_o$), where the option finishes (β_o), and what to do in between these two conditions (π_o).

2 STATE-ACTION ABSTRACTIONS

Together, state and action abstractions can distill complex problems into simple ones (Jonsson & Barto, 2001; Ciosek & Silver, 2015; Bai et al., 2016). Our treatment of state-action abstraction is related to generating options from a bisimulation metric (Ferns et al., 2004) as proposed by Castro & Precup (2011), but distinct from state-action homomorphisms, as explored by Ravindran (2003), Taylor et al. (2008) and Majeed & Hutter (2019). We here introduce a novel means of combining state abstractions with options, defined as follows:

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 $s \in \mathcal{S}$:*

$$\mathcal{I}_o(s) \equiv s \in s_\phi, \qquad \beta_o(s) \equiv s \notin s_\phi, \qquad \pi_o \in \Pi_{s_\phi}, \tag{2}$$

where $\Pi_{s_\phi} : \{s \mid \phi(s) = s_\phi\} \rightarrow \mathcal{A}$ is the set of ground policies defined over states in s_ϕ, and $s \in s_\phi$ is shorthand for $s \in \{\phi(s') = s_\phi \mid \forall_{s' \in \mathcal{S}}\}$. 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, this means we define options that initiate in each abstract state and terminate once the option leaves the abstract state. For example, in the classical Four Rooms domain, if the state abstraction 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 option leaves that room. This gives us a powerful formalism for seamlessly combining state abstractions and options.

We henceforth denote $(\phi, \mathcal{O}_\phi)$ as a state abstraction paired with a set of ϕ-relative options. We first show that, similar to Remark 1, any $(\phi, \mathcal{O}_\phi)$ gives rise to an abstract policy over $\mathcal{S}_\phi$ and $\mathcal{O}_\phi$ that *also* induces a unique policy in the original MDP (over the entire state space). All proofs are presented in the appendix.

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

This theorem 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 $\mathcal{A}$. This process is visualized in Figure 1b. 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, $\Pi^{\Downarrow}_{\phi,\mathcal{O}_\phi}$. The value loss of $\phi, \mathcal{O}_\phi$ is 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 suboptimality achievable:*

$$L(\phi, \mathcal{O}_\phi) := \min_{\pi_{\phi,\mathcal{O}_\phi} \in \Pi_{\phi,\mathcal{O}_\phi}} \left|\left| V^* - V^{\pi^{\Downarrow}_{\phi,\mathcal{O}_\phi}} \right|\right|_\infty . \tag{3}$$

To characterize the loss of various options, we require a final definition that clarifies what is meant by an option class. We adopt a new formalism that characterizes sets of options as containing *representative* options, defined as follows.

Definition 6 (Option Class): *Let $\mathcal{O}^{all}_\phi$ denote the set of all possible ϕ-relative options for a given ϕ. For every s_ϕ, consider a two-place predicate on options of this set, $p_{s_\phi} : \mathcal{O}^{all}_\phi \times \mathcal{O}^{all}_\phi \to \{0, 1\}$. A set of ϕ-relative options is said to belong to the class defined by p_{s_ϕ}, which we denote $\mathcal{O}_{\phi,p}$, if and only if:*

$$\forall_{s_\phi \in \mathcal{S}_\phi} \; \forall_{o_1 \in \mathcal{O}^{all}_\phi} \; \exists_{o_2 \in \mathcal{O}_{\phi,p}} : p_{s_\phi}(o_1, o_2). \tag{4}$$

Intuitively, a class of options consists of choosing a small set of *representative* options from the set of all possible options, and treating those representative options as the set to reason with, $\mathcal{O}_{\phi,p}$. The trick is to choose the representative options appropriately. The predicate defines what counts as a representative option: if p_{s_ϕ} is true of a pair (o_1, o_2), then o_1 is said to be representative of o_2, and vice-versa. In the trivial case, the predicate defines equivalence. If the two options are the same, it is true. In this case, we just recover the set of all options (so every option is its own representative). Instead, we might describe a class of options as those that transition to the same next abstract state from the given s_ϕ; then, we need only retain *one* such option to adhere to this class. Shortly, we will define two classes that possess desirable theoretical properties.

With our definitions in place, we now pose the central question of this work:

> **Central Question:** Are there classes of options that, when paired with well-behaved state abstractions, yield a relatively small $L(\phi, \mathcal{O}_\phi)$?

Our main result answers this question in the affirmative; the following two option classes preserve near-optimality. The option classes we introduce guarantee ε closeness of values or models, building upon state abstraction classes from prior work (Dean & Givan, 1997; Li et al., 2006; Jiang et al., 2015; Abel et al., 2016). More concretely:

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

$$p_{s_\phi}(o_1, o_2) \equiv \max_{s \in s_\phi} |Q^*_{s_\phi}(s, o_1) - Q^*_{s_\phi}(s, o_2)| \leq \varepsilon_Q, \text{where:} \tag{5}$$

$$Q^*_{s_\phi}(s, o) := R(s, \pi_o(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s' \mid s, \pi_o(s)) \left(\mathbb{1}(s' \in s_\phi) Q^*_{s_\phi}(s', o) + \mathbb{1}(s' \notin s_\phi) V^*(s') \right). \tag{6}$$

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

$$p_{s_\phi}(o_1, o_2) \equiv \left|\left|T^{s'}_{s,o_1} - T^{s'}_{s,o_2}\right|\right|_\infty \leq \varepsilon_T \; AND \; ||R_{s,o_1} - R_{s,o_2}||_\infty \leq \varepsilon_R, \text{where:} \tag{7}$$

$R_{s,o}$ and $T^{s'}_{s,o}$ are shorthand for the reward model and multi-time model of Sutton et al. (1999).

Our main result establishes the bounded value loss of these two classes.

Theorem 2. *(Main Result) For any ϕ, the two introduced classes of ϕ-relative options satisfy:*

$$L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \frac{\varepsilon_Q}{1-\gamma}, \qquad L(\phi, \mathcal{O}_{\phi,M_\varepsilon}) \leq \frac{\varepsilon_R + |\mathcal{S}|\varepsilon_T \text{VMAX}}{1-\gamma}. \tag{8}$$

3 HIERARCHIES

We next highlight how the prescribed combination of state abstraction and options can underlie hierarchical RL (Dayan & Hinton, 1993; Parr & Russell, 1998; Dietterich, 2000; Barto & Mahadevan, 2003; Jong & Stone, 2008; Bai & Russell, 2017; Konidaris et al., 2018; Nachum et al., 2019). Specifically, this section presents an extension of Theorem 2 applied 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 (a table summarizing our notation is presented in the appendix). We first define a *hierarchy* as n sets of $(\phi, \mathcal{O}_\phi)$ pairs.

Definition 7 ($(\phi, \mathcal{O}_\phi)$-Hierarchy): *A $(\phi, \mathcal{O}_\phi)$-Hierarchy, denoted H_n, is a list of n state abstractions, $\phi^{(n)}$, and a list of n sets of ϕ-relative options, $\mathcal{O}^{(n)}_\phi$, where 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 $\phi^i(s) = \phi_i(\dots\phi_1(s))$, 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:

$$V^\pi_i(s) := V^\pi_i\left(\phi^i(s)\right) = \max_{o \in \mathcal{O}_i}\left(R_i\left(s_i, o\right) + \sum_{s' \in \mathcal{S}_i} T_i\left(s' \mid s_i, o\right) V^\pi_i(s')\right), \quad \text{where:} \tag{9}$$

$$R_i(s_i, o) := \sum_{s_{i-1} \in s_i} w_i(s_{i-1}) R_{s_{i-1},o}, \quad T_i(s'_i \mid s_i, o) := \sum_{s_{i-1} \in s_i} \sum_{s'_{i-1} \in \mathcal{S}_{i-1}} w_i(s_{i-1}) T^{s'_{i-1}}_{s_{i-1},o},$$

where again $R_{s,o}$ and $T^{s'}_{s,o}$ 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 (Li et al., 2006). Note that V_0 is the ground value function, which we refer to as V for simplicity.

3.1 HIERARCHY ANALYSIS

Our aim is to generalize Theorem 2 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:

Lemma 1. *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. The second key insight is that the same notion of value loss from Definition 2 can be extended to hierarchies, H_n.

Definition 8 (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:*

$$L(H_n) := \min_{\pi_n \in \Pi_n} \left|\left| V^* - V^{\pi_n^{\Downarrow}} \right|\right|_\infty . \tag{10}$$

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 $\mathcal{O}_{\phi,i}$, its value for all states s, when grounded one level down, is similar:*

$$\max_{s \in \mathcal{S}} \left| V_{i-1}^{\pi_i^{\downarrow}} \left(\phi^{i-1}(s) \right) - V_i^{\pi_i} \left(\phi^i(s) \right) \right| \leq \kappa \tag{11}$$

Assumption 2. *Subsequent levels of the hierarchy can represent policies similar in value to 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 ℓ such that:*

$$\min_{\pi_{i+1}^{\downarrow} \in \Pi_{i+1}^{\downarrow}} \left|\left| V_i^{\pi_i^{\diamond}} - V_i^{\pi_{i+1}^{\downarrow}} \right|\right|_\infty \leq \ell. \tag{12}$$

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 2. Regardless, these two assumptions (along with Theorem 2) 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 $L(\phi, \mathcal{O}) \leq \varepsilon_{\mathcal{O}}$. Then, under Assumptions 1 and 2, by repeated application of A_ϕ and $A_{\mathcal{O}_\phi}$, we can construct a hierarchy of depth n such that*

$$L(H_n) = n(\kappa + \ell), \tag{13}$$

where ℓ is some upper bound on $\varepsilon_{\mathcal{O}}$ (and is the same value that appears in Assumption 2).

4 Discussion

We introduce ϕ-relative options, a simple but expressive formalism for combining state abstractions with options. Notably, this method builds options *from* a ϕ function. Using Theorem 1, we prove that any deterministic policy over abstract state and ϕ-relative options induces a single unique policy in the original MDP. This lets us then define the quantity $L(\phi, \mathcal{O}_\phi)$, a coherent notion of value loss extended to capture near-optimality of joint state-action abstractions. We introduce two option classes that trim the space of options down to a smaller representative set. Our main result proves that these two option classes preserve near-optimality in any MDP. We further show that by a simple construction, we can form hierarchies out of ϕ-relative options that also preserve near-optimality. We take these results to serve as a concrete path toward principled abstraction discovery and use.

We are next interested in using insights offered by the analysis presented here to develop reinforcement learning algorithms to find and exploit powerful abstractions that are guaranteed to preserve high quality decision making. To this end, our core direction for future work is to develop a practical option discovery algorithm that 1) offers synergy with state abstraction, and 2) is guaranteed to retain near-optimal behavior. Additionally, we are interested in providing support for both Assumption 1 and 2, as we suspect both are in fact true for many constructions of hierarchies.

References

David Abel, D. Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *ICML*, pp. 2915–2923, 2016.

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

David Andre and Stuart J Russell. State abstraction for programmable reinforcement learning agents. In *AAAI*, pp. 119–125, 2002.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *AAAI*, 2017.

Aijun Bai and Stuart Russell. Efficient reinforcement learning with hierarchies of machines by leveraging internal transitions. In *IJCAI*, 2017.

Aijun Bai, Siddharth Srivastava, and Stuart J Russell. Markovian state and action abstractions for MDPs via hierarchical MCTS. In *IJCAI*, pp. 3029–3039, 2016.

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

Emma Brunskill and Lihong Li. PAC-inspired option discovery in lifelong reinforcement learning. In *ICML*, pp. 316–324, 2014.

Pablo Samuel Castro and Doina Precup. Automatic construction of temporally extended actions for MDPs using bisimulation metrics. In *EWRL*, 2011.

Kamil Ciosek and David Silver. Value iteration with options and state aggregation. *arXiv:1501.03959*, 2015.

Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. In *NeurIPS*, pp. 271–278, 1993.

Thomas Dean and Robert Givan. Model minimization in Markov decision processes. In *AAAI*, 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*, 2000.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *UAI*, 2004.

Ronan Fruit and Alessandro Lazaric. Exploration–exploitation in MDPs with options. *AISTATS*, 2017.

Jesse Hostetler, Alan Fern, and Tom Dietterich. State aggregation in MCTS. In *AAAI*, 2014.

Marcus Hutter. Extreme state aggregation beyond MDPs. In *International Conference on Algorithmic Learning Theory*, pp. 185–199. Springer, 2014.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *ICML*, pp. 179–188, 2015.

Nicholas K Jong and Peter Stone. State abstraction discovery from irrelevant state variables. In *IJCAI*, pp. 752–757, 2005.

Nicholas K Jong and Peter Stone. Hierarchical model-based reinforcement learning: R-max+ MAXQ. In *ICML*, pp. 432–439, 2008.

Anders Jonsson and Andrew G Barto. Automated state abstraction for options using the U-tree algorithm. In *NeurIPS*, pp. 1054–1060, 2001.

George Konidaris and Andrew G Barto. Building portable options: Skill transfer in reinforcement learning. In *IJCAI*, 2007.

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.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. In *ISAIM*, 2006.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. A Laplacian framework for option discovery in reinforcement learning. In *ICML*, 2018.

Sultan Javed Majeed and Marcus Hutter. Performance guarantees for homomorphisms beyond Markov decision processes. *AAAI*, 2019.

Timothy Mann and Shie Mannor. Scaling up approximate value iteration with options: Better policies with fewer iterations. In *ICML*, pp. 127–135, 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. *ICLR*, 2019.

Maillard Odalric-Ambrym, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *ICML*, 2013.

Ronald Parr and Stuart J Russell. Reinforcement learning with hierarchies of machines. In *NeurIPS*, pp. 1043–1049, 1998.

Balaraman Ravindran. SMDP homomorphisms: An algebraic approach to abstraction in semi Markov decision processes. 2003.

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 *NeurIPS*, 2008.

Saket Tiwari and Philip S Thomas. Natural option critic. *AAAI*, 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 *IJCAI*, 2015.

VALUE PRESERVING STATE-ACTION ABSTRACTIONS (APPENDIX)

David Abel[1], **Nathan Umbanhowar**[1], **Khimya Khetarpal**[2], **Dilip Arumugam**[3], **Doina Precup**[2], **Michael L. Littman**[1]
{david_abel,umbanhowar}@brown.edu, khimya.khetarpal@mail.mcgill.ca, dilip@cs.stanford.edu, dprecup@cs.mcgill.ca, mlittman@cs.brown.edu
[1]Brown University, USA
[2]McGill University, CA
[3]Stanford University, USA

1 PROOFS

We here present proofs of each introduced result and Table 1 summarizing notation.

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

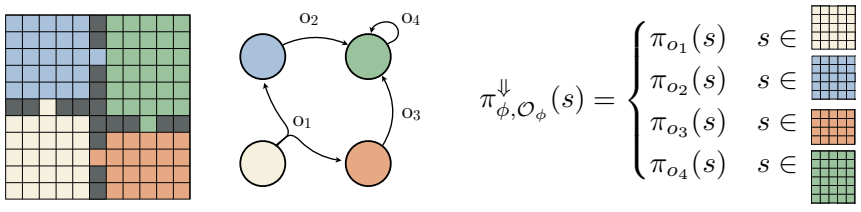
Proof. Consider an arbitrary deterministic policy $\pi_{\phi,\mathcal{O}_\phi}$. By definition, this policy assigns one option to each abstract state. Let $\mathcal{O}_\pi$ denote the set of options this policy assigns.

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 $\pi^{\Downarrow}_{\phi,\mathcal{O}_\phi}$ 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 $\phi(s)$:

$$\pi^{\Downarrow}_{\phi,\mathcal{O}_\phi}(s) = \pi_{o_{\phi(s)}}(s) \tag{16}$$

This construction is visualized in Figure 2. □



(a) Assignment of options to abstract states via $\pi_{\phi,\mathcal{O}_\phi}$.

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

Figure 2: The process of inducing a grounded policy $\pi^{\Downarrow}_{\phi,\mathcal{O}_\phi}$ from $\pi_{\phi,\mathcal{O}_\phi}$.

.....

Theorem 2. *(Main Result) For any ϕ such that $L(\phi) \leq \varepsilon_\phi$, the two introduced classes of ϕ-relative options satisfy:*

$$L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \frac{\varepsilon_Q}{1-\gamma}, \qquad L(\phi, \mathcal{O}_{\phi,M_\varepsilon}) \leq \frac{\varepsilon_R + |\mathcal{S}|\varepsilon_T \text{VMAX}}{1-\gamma}. \tag{17}$$

ϕ	A state abstraction function.
$\mathcal{O}_\phi$	A set of ϕ-relative options.
$\pi_{\phi,\mathcal{O}_\phi}$	A policy that maps each abstract state to an option.
$\pi^{\Downarrow}_{\phi,\mathcal{O}_\phi}$	A policy over $\mathcal{S}$ and $\mathcal{A}$, induced by $\pi_{\phi,\mathcal{O}_\phi}$.
H_n	A hierarchy of depth n, denoting $(\phi^{(n)}, \mathcal{O}_\phi^{(n)})$.
$\phi^{(n)}$	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, $\phi_i(\dots \phi_1(s))$.
$\mathcal{S}_{\phi,i}$	The i-th abstract state space.
V_i^π	The value function of level i policy π defined according to R_i, T_i, $\mathcal{O}_{\phi,i}$, $\mathcal{S}_{\phi,i}$.
$\mathcal{O}_{\phi,i}$	The options available at level i, with each option 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
$\pi_i^{\downarrow}$	A policy over $\mathcal{S}_{\phi,i-1}$ and $\mathcal{O}_{\phi,i-1}$, induced by π_i.
$\pi_i^{\Downarrow}$	A policy over $\mathcal{S}$ and $\mathcal{A}$, induced by π_i.

Table 1: Notation

We prove this claim using two separate proofs, the first targets the $\mathcal{O}_{\phi,Q^*_\varepsilon}$ class of options, and the second, $\mathcal{O}_{\phi,M_\varepsilon}$.

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

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

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

$$o^*_{s_\phi} := \forall_{s\in\mathcal{S}} : \langle \mathcal{I}_{o^*}(s) \equiv \phi(s) = s_\phi, \tag{18}$$

$$\beta(s) \equiv \phi(s) \neq s_\phi, \tag{19}$$

$$\pi(s) = \pi^*(s) \rangle. \tag{20}$$

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 6).

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

$$\forall_{s_\phi\in\mathcal{S}_\phi, s\in s_\phi} : Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \leq \varepsilon_Q. \tag{21}$$

Note that $\mu_{\phi,\mathcal{O}_\phi}(s_\phi)$ outputs an option. As in Equation 21, we henceforth denote $s_\phi = \phi(s)$ and correspondingly $s'_\phi = \phi(s')$.

Then it must be the case that

$$L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \max_{s\in\mathcal{S}} V^*(s) - V^{\mu^{\Downarrow}_{\phi,\mathcal{O}_\phi}}(s). \tag{22}$$

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

$$\lim_{t\to\infty} Q^*_t(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) = V^{\mu^{\Downarrow}_{\phi,\mathcal{O}_\phi}}(s), \tag{23}$$

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

We next show by induction on t that

$$\max_{s\in\mathcal{S}} V^*(s) - V^{\mu^{\Downarrow}_{\phi,\mathcal{O}_\phi}}(s) = \max_{s\in\mathcal{S}} \lim_{t\to\infty} V^*(s) - Q_t^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \leq \frac{\varepsilon_Q}{1-\gamma}. \tag{24}$$

In particular, we wish to show that

$$\forall_{t\in\mathbb{N}} : \max_{s\in\mathcal{S}} V^*(s) - Q_t^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{t} \varepsilon_Q \gamma^i. \tag{25}$$

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

$$Q_0^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) = Q_{s_\phi}^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)), \tag{26}$$

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

$$\begin{aligned}
\max_{s\in\mathcal{S}} V^*(s) - Q_0^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) &= \max_{s\in\mathcal{S}} V^*(s) - Q_{s_\phi}^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)), && (27)\\
&= \max_{s\in\mathcal{S}} Q_{s_\phi}^*(s, o_{s_\phi}^*) - Q_{s_\phi}^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)), && (28)\\
&\leq \varepsilon_Q, && (29)\\
&= \sum_{i=0}^{0} \varepsilon_Q \gamma^0, && (30)
\end{aligned}$$

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

(Inductive Case)
We assume as the inductive hypothesis that

$$\max_{s\in\mathcal{S}} V^*(s) - Q_k^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{k} \varepsilon_Q \gamma^i, \tag{31}$$

and want to show that

$$\max_{s\in\mathcal{S}} V^*(s) - Q_{k+1}^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i. \tag{32}$$

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

$$\begin{aligned}
&V^*(s) - Q_{k+1}^*(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) && (33)\\
&= V^*(s) - \left(R_o(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) + \sum_{s'\in\mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) Q_k^*(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)) \right) && (34)\\
&= V^*(s) - R_o(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) - \sum_{s'\in\mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) Q_k^*(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)) && (35)
\end{aligned}$$

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

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

$$= \quad V^*(s) - R_o(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) - \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))V^*(s') \tag{36}$$

$$+ \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))V^*(s') - \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)) \tag{37}$$

$$= \quad V^*(s) - Q^*_{s_\phi}(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \tag{38}$$

$$+ \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)\right] \tag{39}$$

$$= \quad Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \tag{40}$$

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

$$\leq \quad \varepsilon_Q + \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)\right], \tag{42}$$

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

$$= \quad \varepsilon_Q + \sum_{s' \in \mathcal{S}} \sum_{n=1}^{\infty} \mathbb{P}(s', n|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n \left[V^*(s') - Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)\right] \tag{43}$$

$$\leq \quad \varepsilon_Q + \sum_{s' \in \mathcal{S}} \sum_{n=1}^{\infty} \mathbb{P}(s', n|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n \sum_{i=0}^{k} \varepsilon_Q \gamma^i, \tag{44}$$

$$\tag{45}$$

by the inductive hypothesis. Then:

$$= \quad \varepsilon_Q + \gamma \sum_{s' \in \mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}(s', n+1|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n \sum_{i=0}^{k} \varepsilon_Q \gamma^i \tag{46}$$

$$= \quad \varepsilon_Q + \gamma \sum_{i=0}^{k} \varepsilon_Q \gamma^i \sum_{s' \in \mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}(s', n+1|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n \tag{47}$$

$$\leq \quad \varepsilon_Q + \gamma \sum_{i=0}^{k} \varepsilon_Q \gamma^i \cdot 1 \tag{48}$$

$$= \quad \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i, \tag{49}$$

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

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

$$\max_{s \in \mathcal{S}} V^*(s) - Q^*_{k+1}(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i, \tag{50}$$

as desired.

It follows by induction that

$$\forall_{t \in \mathbb{N}} : \max_{s \in \mathcal{S}} V^*(s) - Q^*_t(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \leq \sum_{i=0}^{t} \varepsilon_Q \gamma^i. \tag{51}$$

Therefore,

$$L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) \leq \max_{s\in\mathcal{S}} V^*(s) - V^{\mu^{\Downarrow}_{\phi,\mathcal{O}_\phi}}(s) \tag{52}$$

$$= \max_{s\in\mathcal{S}} \lim_{t\to\infty} V^*(s) - Q^*_t(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \tag{53}$$

$$\leq \lim_{t\to\infty} \sum_{i=0}^{t} \varepsilon_Q \gamma^i \tag{54}$$

$$= \frac{\varepsilon_Q}{1-\gamma}, \tag{55}$$

which completes the proof.

□

.....

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

Fix $s \in \mathcal{S}$. Let $s_\phi = \phi(s)$. Consider any ϕ-relative option o_1 that initiates in s_ϕ. Then by the M_ε predicate, there exists an option $o_2 \in \mathcal{O}_\phi$ such that

$$||T^{s'}_{s,o_1} - T^{s'}_{s,o_2}||_\infty \leq \varepsilon_T \; AND \; ||R_{s,o_1} - R_{s,o_2}||_\infty \leq \varepsilon_R. \tag{56}$$

Now, we consider the difference in optimal Q-values between o_1 and o_2. We first have that:

$$\begin{aligned} Q^*_{s_\phi}(s, o_1) &= R(s, \pi_{o_1}(s)) + \gamma \sum_{s'\in\mathcal{S}} T(s' \mid s, \pi_{o_1}(s)) \left(\mathbb{1}(s' \in s_\phi) Q^*_{s_\phi}(s', o_1) + \mathbb{1}(s' \notin s_\phi) V^*(s') \right) \\ &= R_o(s, o_1) + \sum_{s'\in\mathcal{S}} T_o(s'|s, o_1) V^*(s'). \end{aligned} \tag{57}$$

By symmetry,

$$Q^*_{s_\phi}(s, o_2) = R_o(s, o_2) + \sum_{s'\in\mathcal{S}} T_o(s'|s, o_2) V^*(s'). \tag{58}$$

Therefore,

$$\begin{aligned} |Q^*_{s_\phi}(s, o_1) - Q^*_{s_\phi}(s, o_2)| &= |R_o(s, o_1) - R_o(s, o_2) + \sum_{s'\in\mathcal{S}} T_o(s'|s, o_1) V^*(s') - \\ &\qquad \sum_{s'\in\mathcal{S}} T_o(s'|s, o_2) V^*(s')| \\ &\leq |R_o(s, o_1) - R_o(s, o_2)| + |\sum_{s'\in\mathcal{S}} \left(T_o(s'|s, o_1) - T_o(s'|s, o_2) \right) V^*(s')| \\ &\leq |R_o(s, o_1) - R_o(s, o_2)| + \sum_{s'\in\mathcal{S}} |T_o(s'|s, o_1) - T_o(s'|s, o_2)||V^*(s')| \\ &\leq \varepsilon_R + |\mathcal{S}|\varepsilon_T \text{VMAX}, \end{aligned} \tag{59}$$

by the model similarity assumption. We have now shown that options with similar models have similar Q-values with $\varepsilon_Q = \varepsilon_R + |\mathcal{S}|\varepsilon_T \text{VMAX}$. Therefore, by the previous result,

$$L(\phi, O_{\phi,M_\varepsilon}) \leq \frac{\varepsilon_R + |\mathcal{S}|\varepsilon_T \text{VMAX}}{1-\gamma}. \tag{60}$$

□

.....

Lemma 1. *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 $\pi_i^{\Downarrow}$.*

Proof. The result follows from an identical strategy to the proof of Theorem 1. □

.....

Theorem 3. *Consider two algorithms:*

1. *A_ϕ: given an MDP M, outputs a ϕ.*
2. *$A_{\mathcal{O}_\phi}$: given M and a ϕ, outputs a set of options $\mathcal{O}$ such that $L(\phi, \mathcal{O}) \leq \varepsilon_{\mathcal{O}}$.*

Then, under Assumptions 1 and 2, by repeated application of A_ϕ and $A_{\mathcal{O}_\phi}$, we can construct a hierarchy of depth n such that

$$L(H_n) = n(\kappa + \ell), \tag{61}$$

where ℓ is some upper bound on $\varepsilon_\phi + \varepsilon_{\mathcal{O}}$ (and is the same value that appears in Assumption 2).

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 2: $$\min_{\pi_1 \in \Pi_1} ||V_0^* - V_0^{\pi_1^\downarrow}||_\infty \leq \ell \tag{62}$$

By Assumption 1: $$\forall_{\pi_1 \in \Pi_1} : ||V_0^{\pi_1^\downarrow} - V_1^{\pi_1}||_\infty \leq \kappa \tag{63}$$

Letting $\pi_1^\diamond = \arg\min_{\pi_1 \in \Pi_1} ||V_0^* - V_0^{\pi_1^\downarrow}||_\infty$, by Assumption 2: $$\min_{\pi_2^\downarrow \in \Pi_2^\downarrow} ||V_1^{\pi_1^\diamond} - V_1^{\pi_2^\downarrow}||_\infty \leq \ell \tag{64}$$

By Assumption 1 $$\forall_{\pi_2^\downarrow \in \Pi_2^\downarrow} : ||V_1^{\pi_2^\downarrow} - V_0^{\pi_2^\Downarrow}||_\infty \leq \kappa \tag{65}$$

Therefore, by the triangle inequality:

$$\min_{\pi_2 \in \Pi_2} ||V_0^* - V_0^{\pi_2^\Downarrow}||_\infty \leq 2\kappa + 2\ell. \tag{66}$$

□

REFERENCES

David Abel, D. Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *ICML*, pp. 2915–2923, 2016.

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

David Andre and Stuart J Russell. State abstraction for programmable reinforcement learning agents. In *AAAI*, pp. 119–125, 2002.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *AAAI*, 2017.

Aijun Bai and Stuart Russell. Efficient reinforcement learning with hierarchies of machines by leveraging internal transitions. In *IJCAI*, 2017.

Aijun Bai, Siddharth Srivastava, and Stuart J Russell. Markovian state and action abstractions for MDPs via hierarchical MCTS. In *IJCAI*, pp. 3029–3039, 2016.

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

Emma Brunskill and Lihong Li. PAC-inspired option discovery in lifelong reinforcement learning. In *ICML*, pp. 316–324, 2014.

Pablo Samuel Castro and Doina Precup. Automatic construction of temporally extended actions for MDPs using bisimulation metrics. In *EWRL*, 2011.

Kamil Ciosek and David Silver. Value iteration with options and state aggregation. *arXiv:1501.03959*, 2015.

Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. In *NeurIPS*, pp. 271–278, 1993.

Thomas Dean and Robert Givan. Model minimization in Markov decision processes. In *AAAI*, 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*, 2000.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *UAI*, 2004.

Ronan Fruit and Alessandro Lazaric. Exploration–exploitation in MDPs with options. *AISTATS*, 2017.

Jesse Hostetler, Alan Fern, and Tom Dietterich. State aggregation in MCTS. In *AAAI*, 2014.

Marcus Hutter. Extreme state aggregation beyond MDPs. In *International Conference on Algorithmic Learning Theory*, pp. 185–199. Springer, 2014.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *ICML*, pp. 179–188, 2015.

Nicholas K Jong and Peter Stone. State abstraction discovery from irrelevant state variables. In *IJCAI*, pp. 752–757, 2005.

Nicholas K Jong and Peter Stone. Hierarchical model-based reinforcement learning: R-max+ MAXQ. In *ICML*, pp. 432–439, 2008.

Anders Jonsson and Andrew G Barto. Automated state abstraction for options using the U-tree algorithm. In *NeurIPS*, pp. 1054–1060, 2001.

George Konidaris and Andrew G Barto. Building portable options: Skill transfer in reinforcement learning. In *IJCAI*, 2007.

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.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. In *ISAIM*, 2006.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. A Laplacian framework for option discovery in reinforcement learning. In *ICML*, 2018.

Sultan Javed Majeed and Marcus Hutter. Performance guarantees for homomorphisms beyond Markov decision processes. *AAAI*, 2019.

Timothy Mann and Shie Mannor. Scaling up approximate value iteration with options: Better policies with fewer iterations. In *ICML*, pp. 127–135, 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. *ICLR*, 2019.

Maillard Odalric-Ambrym, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *ICML*, 2013.

Ronald Parr and Stuart J Russell. Reinforcement learning with hierarchies of machines. In *NeurIPS*, pp. 1043–1049, 1998.

Balaraman Ravindran. SMDP homomorphisms: An algebraic approach to abstraction in semi Markov decision processes. 2003.

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 *NeurIPS*, 2008.

Saket Tiwari and Philip S Thomas. Natural option critic. *AAAI*, 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 *IJCAI*, 2015.