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

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.

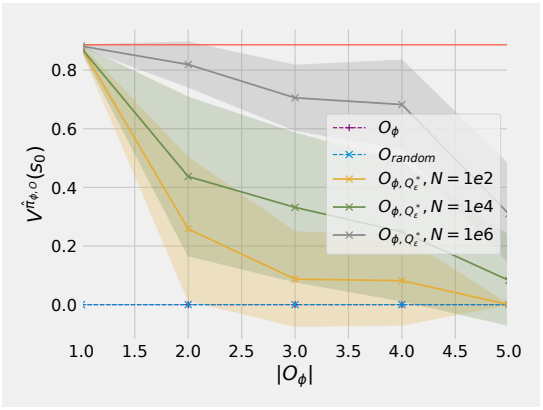
.....

B Experimental Details

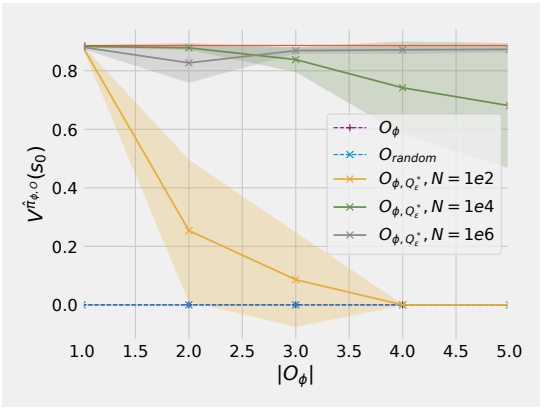
We next provide further detail about the experiment described in Section 3.2.

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.

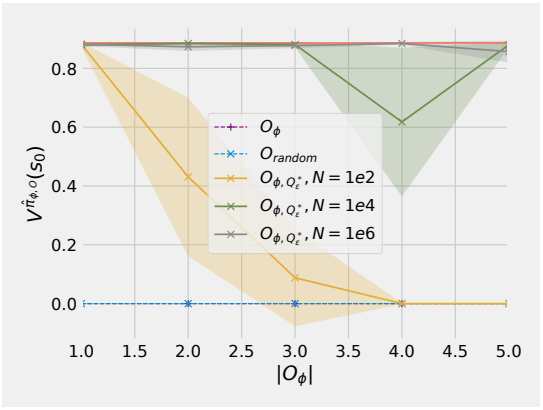
We ran further variations of the experiment with other canonical RL algorithms, including Q-Learning (Watkins and Dayan, 1992), R-Max (Brafman and Tennenholtz, 2002), SARSA (Rummery and Niranjan, 1994), and Delayed Q-Learning (Strehl et al., 2006). Results are presented in Figure 3. Again, we find the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



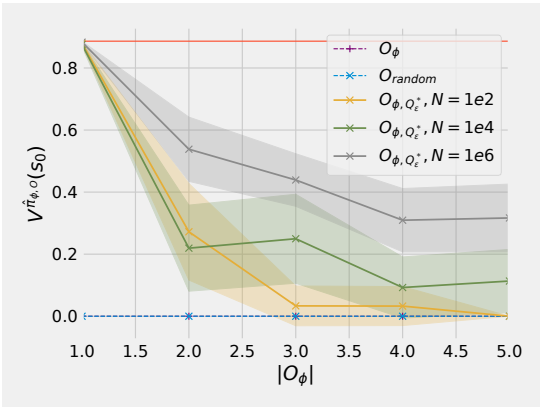
(a) Q-Learning



(b) SARSA



(c) Delayed Q-Learning



(d) R-Max

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

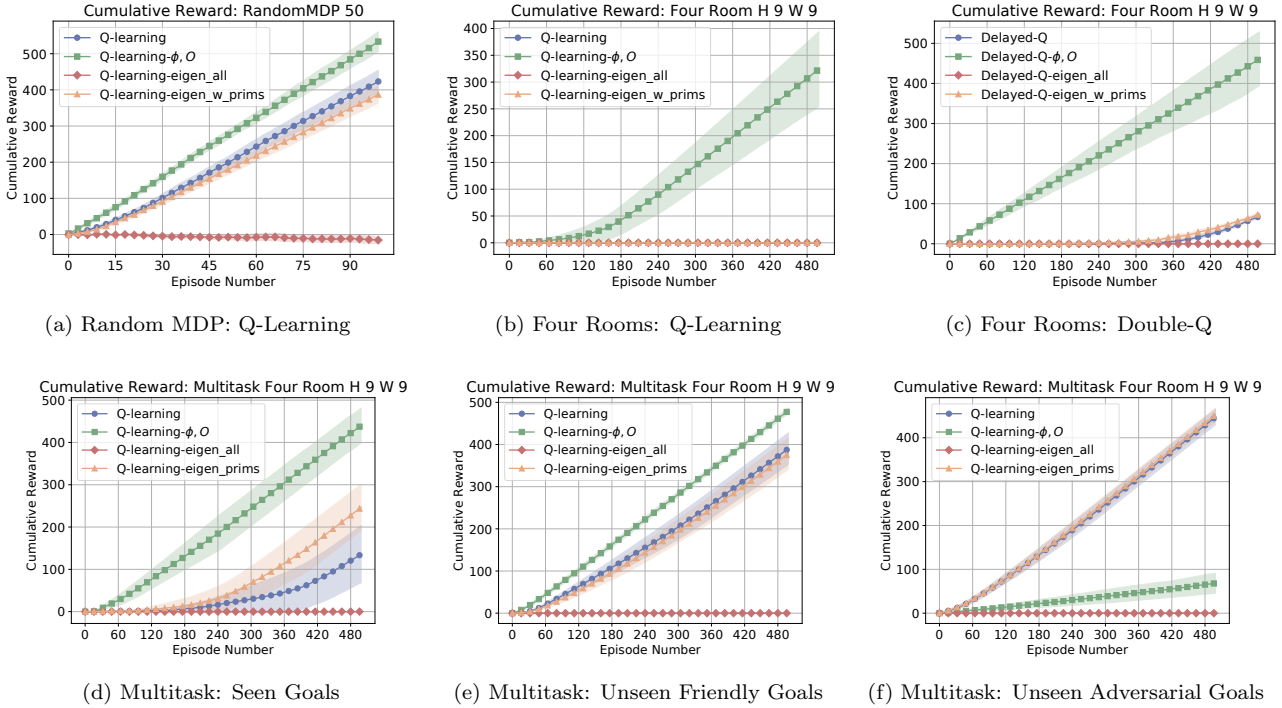
B.1 Experiment: Learning with $\phi, \mathcal{O}_\phi$

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.