Value Preserving State-Action Abstractions: Supplemental Material

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, $\pi_{\phi,\mathcal{O}_\phi} : \mathcal{S}_\phi \rightarrow \mathcal{O}_\phi$, induces a unique Markov policy in the ground MDP, $\pi^{\Downarrow}_{\phi,\mathcal{O}_\phi} : \mathcal{S} \rightarrow \mathcal{A}$. 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.*

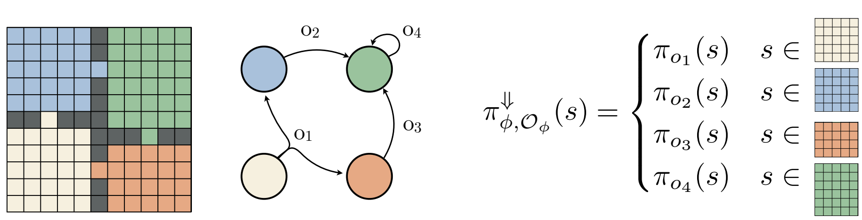
Proof. Consider an arbitrary deterministic policy $\pi_{\phi,\mathcal{O}_\phi}$. By definition, this policy assigns one option to each abstract state. Let $\mathcal{O}_\pi$ denote the set of options this policy assigns.

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

$$
\begin{aligned}
= \quad & V^*(s) - R_o(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) - \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))V^*(s') && (36)\\
& + \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))V^*(s') - \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)) && (37)\\
= \quad & V^*(s) - Q^*_{s_\phi}(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) && (38)\\
& + \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)\right] && (39)\\
= \quad & Q^*_{s_\phi}(s, o^*_{s_\phi}) - Q^*_{s_\phi}(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) && (40)\\
& + \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)\right] && (41)\\
\leq \quad & \varepsilon_Q + \sum_{s' \in \mathcal{S}} T_o(s'|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) \left[V^*(s') - Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)\right], && (42)
\end{aligned}
$$

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

$$
\begin{aligned}
= \quad & \varepsilon_Q + \sum_{s' \in \mathcal{S}} \sum_{n=1}^{\infty} \mathbb{P}(s', n|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n \left[V^*(s') - Q^*_k(s', \mu_{\phi,\mathcal{O}_\phi}(s'_\phi)\right] && (43)\\
\leq \quad & \varepsilon_Q + \sum_{s' \in \mathcal{S}} \sum_{n=1}^{\infty} \mathbb{P}(s', n|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n \sum_{i=0}^{k} \varepsilon_Q \gamma^i, && (44)\\
& && (45)
\end{aligned}
$$

by the inductive hypothesis. Then:

$$
\begin{aligned}
= \quad & \varepsilon_Q + \gamma \sum_{s' \in \mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}(s', n+1|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n \sum_{i=0}^{k} \varepsilon_Q \gamma^i && (46)\\
= \quad & \varepsilon_Q + \gamma \sum_{i=0}^{k} \varepsilon_Q \gamma^i \sum_{s' \in \mathcal{S}} \sum_{n=0}^{\infty} \mathbb{P}(s', n+1|s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi))\gamma^n && (47)\\
\leq \quad & \varepsilon_Q + \gamma \sum_{i=0}^{k} \varepsilon_Q \gamma^i \cdot 1 && (48)\\
= \quad & \sum_{i=0}^{k+1} \varepsilon_Q \gamma^i, && (49)
\end{aligned}
$$

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,

$$
\begin{aligned}
L(\phi, \mathcal{O}_{\phi,Q^*_\varepsilon}) &\leq \max_{s\in\mathcal{S}} V^*(s) - V^{\mu^{\Downarrow}_{\phi,\mathcal{O}_\phi}}(s) && (52)\\
&= \max_{s\in\mathcal{S}} \lim_{t\to\infty} V^*(s) - Q^*_t(s, \mu_{\phi,\mathcal{O}_\phi}(s_\phi)) && (53)\\
&\leq \lim_{t\to\infty} \sum_{i=0}^{t} \varepsilon_Q \gamma^i && (54)\\
&= \frac{\varepsilon_Q}{1-\gamma}, && (55)
\end{aligned}
$$

which completes the proof.

□

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

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:

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

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.