A Proofs of Theorems

We first establish some notation. Let $\mathcal{V} = \{V : \mathcal{S} \to \mathbb{R}_{\geq 0} \mid V \text{ is Lebesgue-measurable and bounded}\}$ denote the set of all concrete value functions and $\tilde{\mathcal{V}} = \{\tilde{V} : \tilde{\mathcal{S}} \to \mathbb{R}_{\geq 0}\}$ denote the set of all abstract value functions. Given $V \in \mathcal{V}$, we denote by $\|V\|_\infty$, the ℓ_∞-norm of V given by $\|V\|_\infty = \sup_{s\in\mathcal{S}} |V(s)|$ and similarly for $\tilde{V} \in \tilde{\mathcal{V}}$, $\|\tilde{V}\|_\infty = \max_{\tilde{s}\in\tilde{\mathcal{S}}} |\tilde{V}(\tilde{s})|$. We use $\mathcal{F}$ to denote the transformation on $\mathcal{V}$ corresponding to (concrete) option value iteration using the set of options $\mathcal{O}$. More precisely, for any $s \in \mathcal{S}$,

$$\mathcal{F}(V)(s) = \max_{o\in\mathcal{O}} Q(V, s, o),$$
$$Q(V, s, o) = R_{\text{opt}}(s, o) + \int_{\mathcal{S}} T_{\text{opt}}(s, o, s')V(s')ds'.$$

We know that $\mathcal{F}$ is a contraction on $\mathcal{V}$ (with respect to the ℓ_∞-norm on $\mathcal{V}$) and hence $\lim_{n\to\infty} \mathcal{F}^n(V)(s) = V^*_{\mathcal{O}}(s)$ for all $s \in \mathcal{S}$ and any initial value function $V \in \mathcal{V}$. Also, for any option policy $\rho : \mathcal{S} \to \mathcal{O}$ we define the corresponding value function V^ρ given by $V^\rho(s) = \lim_{n\to\infty} \mathcal{F}^n_\rho(V)(s)$ where $V \in \mathcal{V}$ is any initial value function and $\mathcal{F}_\rho$ is given by

$$\mathcal{F}_\rho(V)(s) = Q(V, s, \rho(s)).$$

Similarly, for $z \in \{\inf, \sup\}$, let $\tilde{\mathcal{F}}_z : \tilde{\mathcal{V}} \to \tilde{\mathcal{V}}$ denote the transformation corresponding to abstract value iteration—i.e., for any $\tilde{s} \in \tilde{\mathcal{S}}$,

$$\tilde{\mathcal{F}}_z(\tilde{V})(\tilde{s}) = \max_{o\in\mathcal{O}} \tilde{Q}_z(\tilde{V}, \tilde{s}, o),$$
$$\tilde{Q}_z(\tilde{V}, \tilde{s}, o) = \tilde{R}_z(\tilde{s}, o) + \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_z(\tilde{s}, o, \tilde{s}') \cdot \tilde{V}(\tilde{s}').$$

A.1 Proof of Theorem 3.2

We first prove some useful lemmas.

Lemma A.1. *For any finite set $\mathcal{B}$ and two functions $f_1, f_2 : \mathcal{B} \to \mathbb{R}$, if for all $b \in \mathcal{B}$, $|f_1(b) - f_2(b)| \leq \delta$ then $|\max_{b\in\mathcal{B}} f_1(b) - \max_{b\in\mathcal{B}} f_2(b)| \leq \delta$.*

Proof. Let $b_1 = \arg\max_{b\in\mathcal{B}} f_1(b)$ and $b_2 = \arg\max_{b\in\mathcal{B}} f_2(b)$. We need to show that $|f_1(b_1) - f_2(b_2)| \leq \delta$. For the sake of contradiction, suppose $|f_1(b_1) - f_2(b_2)| > \delta$. Then either $f_1(b_1) > f_2(b_2) + \delta$ or $f_2(b_2) > f_1(b_1) + \delta$. Without loss of generality, let us assume $f_1(b_1) > f_2(b_2) + \delta$. Then $f_1(b_1) > f_2(b_1) + \delta$ which implies $|f_1(b_1) - f_2(b_1)| > \delta$, which is a contradiction. □

Lemma A.2. *Given any $\tilde{s} \in \tilde{\mathcal{S}}$ and $o \in \mathcal{O}$,*

$$\sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{inf}(\tilde{s}, o, \tilde{s}') \leq \gamma.$$

Proof. Fix any $s \in \tilde{s}$. Then,

$$\begin{aligned}
\sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{\text{inf}}(\tilde{s}, o, \tilde{s}') &\leq \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}(s, o, \tilde{s}') \\
&= \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \sum_{t=1}^{\infty} \gamma^t P(\tilde{s}', t \mid s, o) \\
&\leq \gamma \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \sum_{t=1}^{\infty} P(\tilde{s}', t \mid s, o) \\
&\leq \gamma
\end{aligned}$$

where the last inequality followed from the fact that the subgoal regions are disjoint. □

Lemma A.3. *For $z \in \{inf, sup\}$,*

$$\sum_{\tilde{s} \in \tilde{\mathcal{S}}} \tilde{T}_z(\tilde{s}, o, \tilde{s}') \leq \gamma + |\tilde{\mathcal{S}}|\varepsilon_T.$$

Proof. The lemma follows from Lemma A.2 and the definition of ε_T. □

A.1.1 Proof of Convergence

We prove that R-AVI converges by showing that abstract value iteration is defined by a contraction mapping. Consider, for any $\tilde{s} \in \tilde{\mathcal{S}}$, $o \in \mathcal{O}$, $\tilde{V}_1, \tilde{V}_2 \in \tilde{\mathcal{V}}$ and $z \in \{\inf, \sup\}$,

$$\begin{aligned}
\left|\tilde{Q}_z(\tilde{V}_1, \tilde{s}, o) - \tilde{Q}_z(\tilde{V}_2, \tilde{s}, o)\right| &= \left| \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{T}_z(\tilde{s}, o, \tilde{s}') \cdot \tilde{V}_1(\tilde{s}') - \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{T}_z(\tilde{s}, o, \tilde{s}') \cdot \tilde{V}_2(\tilde{s}') \right| \\
&= \left| \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{T}_z(\tilde{s}, o, \tilde{s}') \cdot \left(\tilde{V}_1(\tilde{s}') - \tilde{V}_2(\tilde{s}')\right) \right| \\
&\leq \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{T}_z(\tilde{s}, o, \tilde{s}') \cdot \left|\tilde{V}_1(\tilde{s}') - \tilde{V}_2(\tilde{s}')\right| \\
&\leq \|\tilde{V}_1 - \tilde{V}_2\|_\infty \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{T}_z(\tilde{s}, o, \tilde{s}') \\
&\leq (\gamma + |\tilde{\mathcal{S}}|\varepsilon_T)\|\tilde{V}_1 - \tilde{V}_2\|_\infty.
\end{aligned}$$

where the last inequality followed from Lemma A.3. Using Lemma A.1 we have

$$\begin{aligned}
|\tilde{\mathcal{F}}_z(\tilde{V}_1)(\tilde{s}) - \tilde{\mathcal{F}}_z(\tilde{V}_2)(\tilde{s})| &= \left| \max_{o \in \mathcal{O}} \tilde{Q}_z(\tilde{V}_1, \tilde{s}, o) - \max_{o \in \mathcal{O}} \tilde{Q}_z(\tilde{V}_2, \tilde{s}, o) \right| \\
&\leq (\gamma + |\tilde{\mathcal{S}}|\varepsilon_T)\|\tilde{V}_1 - \tilde{V}_2\|_\infty.
\end{aligned}$$

If $\gamma + |\tilde{\mathcal{S}}|\varepsilon_T < 1$, $\tilde{\mathcal{F}}_z$ is a contraction mapping and hence abstract value iteration is guaranteed to converge. □

A.1.2 Proof of Performance Bound

We show the performance bound using the following lemmas. First, we show that the upper and lower values obtained from abstract value iteration bound the value function of the best option policy ρ^* for the set of options $\mathcal{O}$.

Lemma A.4. *Under Assumption 3.1, for all $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$, we have*

$$\tilde{V}^*_{inf}(\tilde{s}) \leq V^*_{\mathcal{O}}(s) \leq \tilde{V}^*_{sup}(\tilde{s}).$$

Proof. We will prove the upper bound. The lower bound follows by a similar argument. Let $V \in \mathcal{V}$ and $\tilde{V} \in \tilde{\mathcal{V}}$ be such that for all $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$, $V(s) \leq \tilde{V}(\tilde{s})$. Suppose $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$. Since for any $o \in \mathcal{O}$,

$\int_{\mathcal{S}} T_{\text{opt}}(s,o,s')\mathbb{1}(s' \in \mathcal{S} \setminus \bar{S})ds' = 0$, we have

$$\begin{aligned}
\mathcal{F}(V)(s) &= \max_{o\in\mathcal{O}} \Big(R_{\text{opt}}(s,o) + \int_{\mathcal{S}} T_{\text{opt}}(s,o,s')V(s')ds'\Big) \\
&= \max_{o\in\mathcal{O}} \Big(R_{\text{opt}}(s,o) + \int_{\bar{S}} T_{\text{opt}}(s,o,s')V(s')ds'\Big) \\
&= \max_{o\in\mathcal{O}} \Big(R_{\text{opt}}(s,o) + \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \int_{\bar{s}'} T_{\text{opt}}(s,o,s')V(s')ds'\Big) \\
&\le \max_{o\in\mathcal{O}} \Big(\tilde{R}_{\text{sup}}(\tilde{s},o) + \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{V}(\tilde{s}') \int_{\bar{s}'} T_{\text{opt}}(s,o,s')ds'\Big) \\
&= \max_{o\in\mathcal{O}} \Big(\tilde{R}_{\text{sup}}(\tilde{s},o) + \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}(s,o,\tilde{s}')\cdot\tilde{V}(\tilde{s}')\Big) \\
&\le \max_{o\in\mathcal{O}} \Big(\tilde{R}_{\text{sup}}(\tilde{s},o) + \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{\text{sup}}(\tilde{s},o,\tilde{s}')\cdot\tilde{V}(\tilde{s}')\Big) \\
&= \tilde{\mathcal{F}}_{\text{sup}}(\tilde{V})(\tilde{s}).
\end{aligned}$$

By induction on n, it follows that $\mathcal{F}^n(V)(s) \le \tilde{\mathcal{F}}^n_{\text{sup}}(\tilde{V})(\tilde{s})$ for all $n \ge 1$. Therefore if V_0 and $\tilde{V}_0$ assign zero to all states and subgoal regions, respectively, we have

$$V^*_{\mathcal{O}}(s) = \lim_{n\to\infty} \mathcal{F}^n(V_0)(s) \le \lim_{n\to\infty} \tilde{\mathcal{F}}^n_{\text{sup}}(\tilde{V}_0)(\tilde{s}) = \tilde{V}^*_{\text{sup}}(\tilde{s}).$$

The claim follows. □

Next, we bound the gap in the upper and lower value functions as a function of the gaps ε_T and ε_R.

Lemma A.5. *Under Assumption 3.1, for all $\tilde{s} \in \tilde{\mathcal{S}}$, we have*

$$\tilde{V}^*_{sup}(\tilde{s}) - \tilde{V}^*_{inf}(\tilde{s}) \le \frac{(1-\gamma)\varepsilon_R + |\tilde{\mathcal{S}}|\varepsilon_T}{(1-\gamma)(1-(\gamma+|\tilde{\mathcal{S}}|\varepsilon_T))}.$$

Proof. Let $\tilde{V}_1, \tilde{V}_2 \in \tilde{\mathcal{V}}$ be abstract value functions such that $\tilde{V}_2(\tilde{s}) \le \min\{(1-\gamma)^{-1}, \tilde{V}_1(\tilde{s})\}$ for all $\tilde{s} \in \tilde{\mathcal{S}}$. Then, for any $\tilde{s} \in \tilde{\mathcal{S}}$ and $o \in \mathcal{O}$,

$$\begin{aligned}
&\tilde{Q}_{\text{sup}}(\tilde{V}_1,\tilde{s},o) - \tilde{Q}_{\text{inf}}(\tilde{V}_2,\tilde{s},o) \\
&\quad= \Big(\tilde{R}_{\text{sup}}(\tilde{s},o) - \tilde{R}_{\text{inf}}(\tilde{s},o)\Big) + \Big(\sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{\text{sup}}(\tilde{s},o,\tilde{s}')\cdot\tilde{V}_1(\tilde{s}') - \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{\text{inf}}(\tilde{s},o,\tilde{s}')\cdot\tilde{V}_2(\tilde{s}')\Big) \\
&\quad\le \varepsilon_R + \Big(\sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{\text{sup}}(\tilde{s},o,\tilde{s}')\cdot\tilde{V}_1(\tilde{s}') - \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \big(\tilde{T}_{\text{sup}}(\tilde{s},o,\tilde{s}') - \varepsilon_T\big)\cdot\tilde{V}_2(\tilde{s}')\Big) \\
&\quad\le \varepsilon_R + \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{\text{sup}}(\tilde{s},o,\tilde{s}')\cdot\big(\tilde{V}_1(\tilde{s}') - \tilde{V}_2(\tilde{s}')\big) + \frac{|\tilde{\mathcal{S}}|\varepsilon_T}{1-\gamma} \\
&\quad\le \varepsilon_R + \|\tilde{V}_1 - \tilde{V}_2\|_\infty \sum_{\tilde{s}'\in\tilde{\mathcal{S}}} \tilde{T}_{\text{sup}}(\tilde{s},o,\tilde{s}') + \frac{|\tilde{\mathcal{S}}|\varepsilon_T}{1-\gamma} \\
&\quad\le \varepsilon_R + (\gamma + |\tilde{\mathcal{S}}|\varepsilon_T)\|\tilde{V}_1 - \tilde{V}_2\|_\infty + \frac{|\tilde{\mathcal{S}}|\varepsilon_T}{1-\gamma}.
\end{aligned}$$

Now, using Lemma A.1 we have

$$|\tilde{\mathcal{F}}_{\text{sup}}(\tilde{V}_1)(\tilde{s}) - \tilde{\mathcal{F}}_{\text{inf}}(\tilde{V}_2)(\tilde{s})| \le \varepsilon_R + (\gamma + |\tilde{\mathcal{S}}|\varepsilon_T)\|\tilde{V}_1 - \tilde{V}_2\|_\infty + \frac{|\tilde{\mathcal{S}}|\varepsilon_T}{1-\gamma}.$$

If we define $\tilde{V}_0$ to be the zero vector, we can show by induction on n that, for all $\tilde{s} \in \tilde{\mathcal{S}}$ and $n \geq 0$, $\tilde{\mathcal{F}}^n_{\text{inf}}(\tilde{V}_0)(\tilde{s}) \leq \min\{(1-\gamma)^{-1}, \tilde{\mathcal{F}}^n_{\text{sup}}(\tilde{V}_0)(\tilde{s})\}$ since the rewards in the underlying MDP are bounded above by 1. Hence, another induction on n gives us, for all $\tilde{s} \in \tilde{\mathcal{S}}$ and $n \geq 0$,

$$\tilde{\mathcal{F}}^n_{\text{sup}}(\tilde{V}_0)(\tilde{s}) - \tilde{\mathcal{F}}^n_{\text{inf}}(\tilde{V}_0)(\tilde{s}) \leq \Big(\varepsilon_R + \frac{|\tilde{\mathcal{S}}|\varepsilon_T}{1-\gamma}\Big) \sum_{k=0}^{n} (\gamma + |\tilde{\mathcal{S}}|\varepsilon_T)^k.$$

Taking limit $n \to \infty$ on both sides gives us the required bound. □

Now, we prove the following lemma.

Lemma A.6. *For any $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$ we have*

$$V^{\tilde{\rho}}(s) \geq \tilde{V}^*_{inf}(\tilde{s}),$$

where $\tilde{\rho}$ is the conservative optimal option policy.

Proof. Let $V \in \mathcal{V}$ be such that for all $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$, $V(s) \geq \tilde{V}^*_{\text{inf}}(\tilde{s})$. Given $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$, we have

$$\begin{aligned}
\mathcal{F}_{\tilde{\rho}}(V)(s) &= R_{\text{opt}}(s, \tilde{\rho}(s)) + \int_{\mathcal{S}} T_{\text{opt}}(s, \tilde{\rho}(s), s') V(s') ds' \\
&\geq \tilde{R}_{\text{inf}}(\tilde{s}, \tilde{\rho}(s)) + \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{V}^*_{\text{inf}}(\tilde{s}') \int_{\tilde{s}'} T_{\text{opt}}(s, \tilde{\rho}(s), s') ds' \\
&\geq \tilde{R}_{\text{inf}}(\tilde{s}, \tilde{\rho}(s)) + \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{T}(s, \tilde{\rho}(s), \tilde{s}') \cdot \tilde{V}^*_{\text{inf}}(\tilde{s}') \\
&\geq \tilde{R}_{\text{inf}}(\tilde{s}, \tilde{\rho}(s)) + \sum_{\tilde{s}' \in \tilde{\mathcal{S}}} \tilde{T}_{\text{inf}}(\tilde{s}, \tilde{\rho}(s), \tilde{s}') \cdot \tilde{V}^*_{\text{inf}}(\tilde{s}') \\
&= \tilde{Q}^*_{\text{inf}}(\tilde{s}, \tilde{\rho}(s)) \\
&= \max_{o \in \mathcal{O}} \tilde{Q}^*_{\text{inf}}(\tilde{s}, o) \\
&= \tilde{V}^*_{\text{inf}}(\tilde{s})
\end{aligned}$$

where the first inequality followed from the fact that $\int_{\mathcal{S}} T_{\text{opt}}(s, o, s') \mathbb{1}(s' \in \mathcal{S} \setminus \bar{S}) ds' = 0$. Now let $V_0 \in \mathcal{V}$ be a value function such that $V_0(s) = \tilde{V}^*_{\text{inf}}(\tilde{s})$ for all $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$. Then we can show by induction on n that, for all $\tilde{s} \in \tilde{\mathcal{S}}$, $s \in \tilde{s}$ and $n \geq 0$, $\mathcal{F}^n_{\tilde{\rho}}(V_0)(s) \geq \tilde{V}^*_{\text{inf}}(\tilde{s})$ and therefore

$$V^{\tilde{\rho}}(s) = \lim_{n \to \infty} \mathcal{F}^n_{\tilde{\rho}}(V_0)(s) \geq \tilde{V}^*_{\text{inf}}(\tilde{s}).$$

The claim follows. □

We are now ready to prove the performance bound in Theorem 3.2. For any $\tilde{s} \in \tilde{\mathcal{S}}$ and $s \in \tilde{s}$, we have

$$\begin{aligned}
V^{\tilde{\rho}}(s) &\geq \tilde{V}^*_{\text{inf}}(\tilde{s}) \\
&= \tilde{V}^*_{\text{sup}}(\tilde{s}) - (\tilde{V}^*_{\text{sup}}(\tilde{s}) - \tilde{V}^*_{\text{inf}}(\tilde{s})) \\
&\geq V^*_{\mathcal{O}}(s) - (\tilde{V}^*_{\text{sup}}(\tilde{s}) - \tilde{V}^*_{\text{inf}}(\tilde{s}))
\end{aligned}$$

where the first inequality followed from Lemma A.6 and the second inequality followed from Lemma A.4. Taking expectation w.r.t. the initial state distribution η_0 and applying Lemma A.5 gives us the required claim. □

A.2 Proof of Theorem 3.4

Note that this theorem relies on additional assumptions, namely, Assumptions 2.2 and 3.3. We first show the following lemma.[5]

[5]Note that $\tilde{V}^*_{\text{sup}}$ is an upper bound on the value function; it may exceed the optimal value.

Lemma A.7. *For all $s_0 \in \tilde{s}_0$, $V^*(s_0) \leq \tilde{V}^*_{sup}(\tilde{s}_0)$.*

Proof. Let π^* be the optimal policy. Given an $s_0 \in \tilde{s}_0$, let $s_0, s_1, \ldots$ be the sequence of states visited when following π^* starting at s_0. If the goal region is not visited, then $V^*(s_0) = 0$ and the lemma holds. Otherwise, let t be the first time when $s_t \in \tilde{s}_g$. Then $V^*(s_0) = \gamma^{t-1}$ and there is a subsequence of indices, $0 = i_0, \ldots, i_k = t$ and a sequence of subgoal region $\tilde{s}_0, \ldots, \tilde{s}_k$ such that for all $0 \leq j \leq k$, $s_{i_j} \in \tilde{s}_j$ and for $j < k$, there is an option $o_j = (\pi(\tilde{s}_j, \tilde{s}_{j+1}), \tilde{s}_j, \beta) \in \mathcal{O}^*$. Let o^*_j denote the modified option $(\pi^*, \tilde{s}_j, \beta)$ where the policy $\pi(\tilde{s}_j, \tilde{s}_{j+1})$ is replaced with π^*. For every $0 \leq j < k$,

$$\begin{aligned}
\gamma^{i_{j+1}-i_j} &= \tilde{T}(s_{i_j}, o^*_j, \tilde{s}_{j+1}) \\
&\leq \tilde{T}_{\text{sup}}(\tilde{s}_j, o^*_j, \tilde{s}_{j+1}) \\
&\leq \max_{\pi} \tilde{T}_{\text{sup}}(\tilde{s}_j, (\pi, \tilde{s}_j, \beta), \tilde{s}_{j+1}) \\
&= \tilde{T}_{\text{sup}}(\tilde{s}_j, o_j, \tilde{s}_{j+1}).
\end{aligned}$$

Since all states in $\tilde{s}_g$ are sink states, $\tilde{V}^*_{\text{sup}}(\tilde{s}_g) = 0$. Furthermore, for any $s \in \bar{S} \setminus \tilde{s}_g$ and any subgoal transition o, $R_{\text{opt}}(s, o) = \gamma^{-1}\tilde{T}(s, o, \tilde{s}_g)$ and hence

$$\begin{aligned}
\tilde{R}_{\text{sup}}(\tilde{s}_{k-1}, o_{k-1}) &= \sup_{s \in \tilde{s}_{k-1}} R_{\text{opt}}(s, o_{k-1}) \\
&= \sup_{s \in \tilde{s}_{k-1}} \gamma^{-1}\tilde{T}(s, o_{k-1}, \tilde{s}_g) \\
&= \gamma^{-1}\tilde{T}_{\text{sup}}(\tilde{s}_{k-1}, o_{k-1}, \tilde{s}_g) \\
&\geq \gamma^{t-i_{k-1}-1}.
\end{aligned}$$

Since $\tilde{R}_{\text{sup}}(\tilde{s}_j, o_j) \geq 0$ for all $0 \leq j < k$, using the definition of $\tilde{V}^*_{\text{sup}}$ and induction on $k - j$ we can show that for all $0 \leq j < k$,

$$\tilde{V}^*_{\text{sup}}(\tilde{s}_j) \geq \tilde{R}_{\text{sup}}(\tilde{s}_{k-1}, o_{k-1}) \prod_{q=j}^{k-2} \tilde{T}_{\text{sup}}(\tilde{s}_q, o_q, \tilde{s}_{q+1}) \geq \gamma^{t-i_j-1}$$

Therefore, $\tilde{V}^*_{\text{sup}}(\tilde{s}_0) \geq \gamma^{t-1} = V^*(s_0)$. □

We are now ready to prove Theorem 3.4. We have

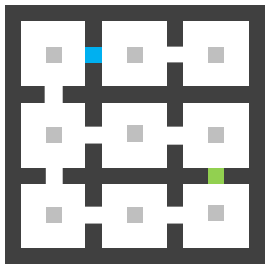
$$\begin{aligned}
J(\pi^*) - J(\pi_{\tilde{\rho}}) &= \mathbb{E}_{s_0 \sim \eta_0}[V^*(s_0) - V^{\tilde{\rho}}(s_0)] \\
&\leq \mathbb{E}_{s_0 \sim \eta_0}[\tilde{V}^*_{\text{sup}}(\tilde{s}_0) - \tilde{V}^*_{\text{inf}}(\tilde{s}_0)] \\
&\leq \frac{(1-\gamma)\varepsilon_R + |\tilde{\mathcal{S}}|\varepsilon_T}{(1-\gamma)(1-(\gamma + |\tilde{\mathcal{S}}|\varepsilon_T))},
\end{aligned}$$

where the first inequality followed from Lemmas A.7 & A.6, and the second inequality followed from Lemma A.5[6]. □

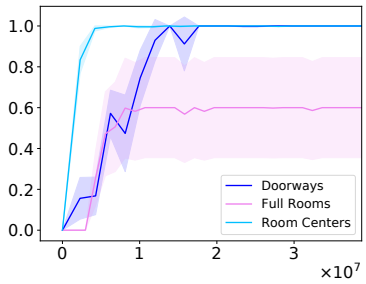
B Experimental Details

Additional Figures. Subgoal regions given by "room centers" in the 9-Rooms environment are visualized in Figure 6 (a). The learning curves for different choices of subgoal regions for the room environments are shown in Figure 6 (b,c) where we plot the probability of reaching the goal as a function of the number of steps taken in the environment; in contrast, the cumulative reward plotted in Figure 3 measures not only the probability of reaching the goal but also the time to reach the goal. In particular, "room centers" can also be used to learn a policy that reaches the goal with an estimated probability of 1, although they do not satisfy the bottleneck assumption. Thus, this choice of subgoal regions only reduces the time to reach the goal, not the probability of reaching the goal.

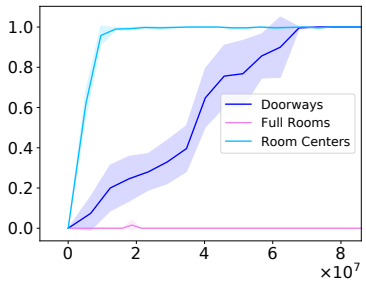
[6]Although we assumed that $T(s, a, s') = p(s' \mid s, a)$ defines a probability density function, it is easy to see that lemmas hold true for the deterministic case as well.



(a) Room Centers

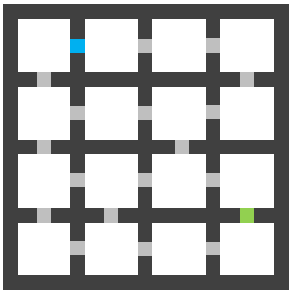


(b) 9-Rooms

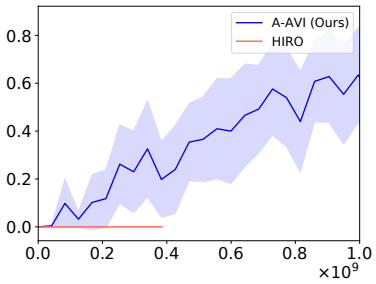


(c) 16-Rooms

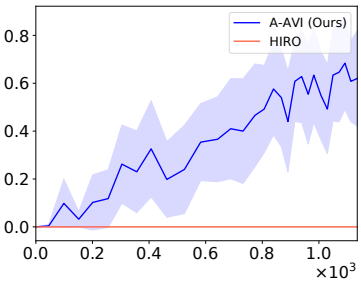
Figure 6: Visualization of room centers as subgoal regions (in gray) and comparison of subgoal regions for room environments; x-axis is number of samples (steps) from the environment, and y-axis is probability of reaching the goal. Results are averaged over 10 executions.



(a) 16-Rooms



(b) Sample Complexity



(c) Learning Time

Figure 7: The 16-Rooms environment and learning curves of A-AVI with randomly generated subgoal regions in 16-Rooms; the plots show the probability of reaching the goal (y-axis) as a function of (b) number of samples (steps) from the environment and (c) time since the beginning of training (in minutes). Results are averaged over 10 executions.

The 16-Rooms environment is visualized in Figure 7 (a). We also trained policies for the 16-Rooms environment using randomly generated subgoal regions. For this environment we used $N = 25$ subgoal regions and $K = 7$ outgoing edges from each subgoal region. As shown in Figure 7 (b,c) we outperform HIRO on this task as well without additional input from the user.

The subgoal regions for AntMaze, AntPush, and AntFall are visualized in Figures 8, 9, and 10, respectively. The red squares are the subgoal regions; in particular, each subgoal region can be described as a constraint $x \in [x_{\min}, x_{\max}] \wedge y \in [y_{\min}, y_{\max}]$, where $(x, y) \in \mathbb{R}^2$ is the position of the center of the ant.

Hyperparameters. For the rooms environment, the subgoal regions are learned using ARS (Mania et al., 2018) (version V2-t) with neural network policies and the following hyperparameters.

- Step-size $\alpha = 0.3$.
- Standard deviation of exploration noise $\nu = 0.05$.
- Number of directions sampled per iteration is 30.
- Number of top performing directions to use $b = 15$.

We retain the parameters of the policies across iterations of A-AVI. In each iteration of A-AVI, we run 300 iterations of ARS for each subgoal transition in parallel. Initially, $\mathcal{D}_{\tilde{s}}$ is taken to be the uniform distribution in a small square in the center of the subgoal region $\tilde{s}$.

Figure 8: Subgoal Regions for AntMaze

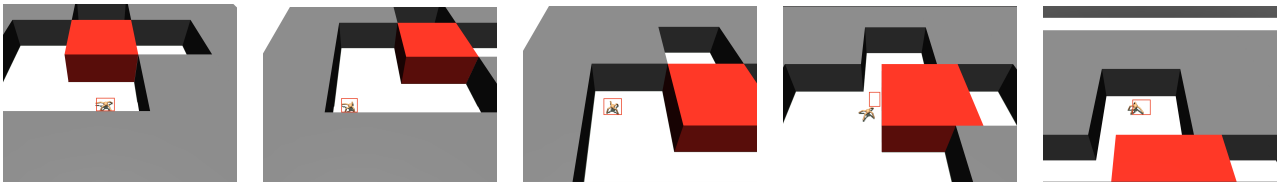


Figure 9: Subgoal Regions for AntPush

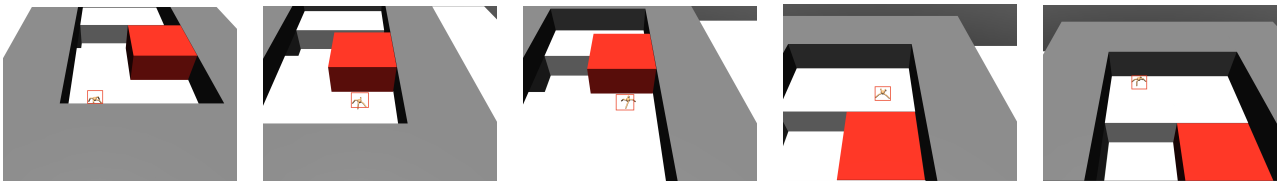


Figure 10: Subgoal Regions for AntFall

For the ant environments, the subgoal transitions are learned using TD3 (Fujimoto et al., 2018); each policy is a fully connected neural network with 300 neurons each and critic architecture is the same as the one in Fujimoto et al. (2018) except that we use 300 neurons for both hidden layers. We use the TFAgents (Guadarrama et al., 2018) implementation of TD3 with the following hyperparameters.

- Discount $\gamma = 0.95$.
- Adam optimizer; actor learning rate 0.0001; critic learning rate 0.001.
- Soft update targets $\tau = 0.005$.
- Replay buffer of size 200000.
- Target update and training step performed every 2 environment steps.
- Exploration using gaussian noise with $\sigma = 0.1$.

We retain the actor and critic networks, target networks, optimizer states and the replay buffers across iterations of A-AVI. In each iteration of A-AVI, we run TD3 for 100000 environment steps for each subgoal transition.