SUPPLEMENTARY MATERIAL: GENERALISATION IN LIFELONG REINFORCEMENT LEARNING THROUGH LOGICAL COMPOSITION

Anonymous authors
Paper under double-blind review

1 PROOFS OF THEORETICAL RESULTS

1.1 BOOLEAN ALGEBRA DEFINITION

Definition 1. *A Boolean algebra is a set $\mathcal{B}$ equipped with the binary operators $\vee$ (disjunction) and $\wedge$ (conjunction), and the unary operator $\neg$ (negation), which satisfies the following Boolean algebra axioms for a, b, c in $\mathcal{B}$:*

(i) Idempotence: $a \wedge a = a \vee a = a$.

(ii) Commutativity: $a \wedge b = b \wedge a$ and $a \vee b = b \vee a$.

(iii) Associativity: $a \wedge (b \wedge c) = (a \wedge b) \wedge c$ and $a \wedge (b \vee c) = (a \vee b) \vee c$.

(iv) Absorption: $a \wedge (a \vee b) = a \vee (a \wedge b) = a$.

(v) Distributivity: $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ and $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$.

(vi) Identity: there exists $\mathbf{0}, \mathbf{1}$ in $\mathcal{B}$ such that

$$\begin{aligned} \mathbf{0} \wedge a &= \mathbf{0} \\ \mathbf{0} \vee a &= a \\ \mathbf{1} \wedge a &= a \\ \mathbf{1} \vee a &= \mathbf{1} \end{aligned}$$

(vii) Complements: for every a in $\mathcal{B}$, there exists an element a' in $\mathcal{B}$ such that $a \wedge a' = \mathbf{0}$ and $a \vee a' = \mathbf{1}$.

1.2 PROOFS FOR PROPOSITION 2

Lemma 1. *Let $\mathcal{M}$ be a set of tasks. Then $(\mathcal{M}, \vee, \wedge, \neg, \mathcal{M}_{MAX}, \mathcal{M}_{MIN})$ is a Boolean algebra.*

Proof. Let $M_1, M_2 \in \mathcal{M}$. We show that $\neg, \vee, \wedge$ satisfy the Boolean properties (i) – (vii).

(i)–(v): These easily follow from the fact that the $\min$ and $\max$ functions satisfy the idempotent, commutative, associative, absorption and distributive laws.

(vi): Let $r_{\mathcal{M}_{MAX} \wedge M_1}$ and r_{M_1} be the reward functions for $\mathcal{M}_{MAX} \wedge M_1$ and M_1 respectively. Then for all (s, a) in $\mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} r_{\mathcal{M}_{MAX} \wedge M_1}(s,a) &= \begin{cases} \min\{r_{\text{MAX}}, r_{M_1}(s,a)\}, & \text{if } s \in \mathcal{G} \\ \min\{r_0(s,a), r_0(s,a)\}, & \text{otherwise.} \end{cases} \\ &= \begin{cases} r_{M_1}(s,a), & \text{if } s \in \mathcal{G} \\ r_0(s,a), & \text{otherwise.} \end{cases} && (r_{M_1}(s,a) \in \{r_{\text{MIN}}, r_{\text{MAX}}\} \text{ for } s \in \mathcal{G}) \\ &= r_{M_1}(s,a). \end{aligned}$$

Thus $\mathcal{M}_{MAX} \wedge M_1 = M_1$. Similarly $\mathcal{M}_{MAX} \vee M_1 = \mathcal{M}_{MAX}$, $\mathcal{M}_{MIN} \wedge M_1 = \mathcal{M}_{MIN}$, and $\mathcal{M}_{MIN} \vee M_1 = M_1$. Hence $\mathcal{M}_{MIN}$ and $\mathcal{M}_{MAX}$ are the universal bounds of $\mathcal{M}$.

(vii): Let $r_{M_1 \wedge \neg M_1}$ be the reward function for $M_1 \wedge \neg M_1$. Then for all (s, a) in $\mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} r_{M_1 \wedge \neg M_1}(s,a) &= \begin{cases} \min\{r_{M_1}(s,a), (r_{\text{MAX}} + r_{\text{MIN}}) - r_{M_1}(s,a)\}, & \text{if } s \in \mathcal{G} \\ \min\{r_0(s,a), (r_0(s,a) + r_0(s,a)) - r_0(s,a)\}, & \text{otherwise.} \end{cases} \\ &= \begin{cases} r_{\text{MIN}}, & \text{if } s \in \mathcal{G} \text{ and } r_{M_1}(s,a) = r_{\text{MAX}} \\ r_{\text{MAX}}, & \text{if } s \in \mathcal{G} \text{ and } r_{M_1}(s,a) = r_{\text{MIN}} \\ r_0(s,a), & \text{otherwise.} \end{cases} \\ &= r_{\mathcal{M}_{MIN}}(s,a). \end{aligned}$$

Thus $M_1 \wedge \neg M_1 = \mathcal{M}_{MIN}$, and similarly $M_1 \vee \neg M_1 = \mathcal{M}_{MAX}$.

□

Lemma 2. *Let $\bar{\mathcal{Q}}^*$ be the set of optimal $\bar{Q}$-value functions for tasks in $\mathcal{M}$. Then $(\bar{\mathcal{Q}}^*, \vee, \wedge, \neg, \bar{Q}^*_{MAX}, \bar{Q}^*_{MIN})$ is a Boolean Algebra.*

Proof. Let $\bar{Q}^*_{M_1}, \bar{Q}^*_{M_2} \in \bar{\mathcal{Q}}^*$ be the optimal $\bar{Q}$-value functions for tasks $M_1, M_2 \in \mathcal{M}$ with reward functions r_{M_1} and r_{M_2}. We show that $\neg, \vee, \wedge$ satisfy the Boolean properties (i) – (vii).

(i)–(v): These follow directly from the properties of the $\min$ and $\max$ functions.

(vi): For all (s, g, a) in $\mathcal{S} \times \mathcal{G} \times \mathcal{A}$,

$$\begin{aligned} (\bar{Q}^*_{MAX} \wedge \bar{Q}^*_{M_1})(s,g,a) &= \min\{\bar{Q}^*_{MAX}(s,g,a), \bar{Q}^*_{M_1}(s,g,a)\} \\ &= \begin{cases} \min\{\bar{Q}^*_{MAX}(s,g,a), \bar{Q}^*_{MAX}(s,g,a)\}, & \text{if } r_{M_1}(g,a') = r_{\text{MAX}}\ \forall a' \in \mathcal{A} \\ \min\{\bar{Q}^*_{MAX}(s,g,a), \bar{Q}^*_{MIN}(s,g,a)\}, & \text{otherwise.} \end{cases} \\ &= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } r_{M_1}(g,a) = r_{\text{MAX}}\ \forall a' \in \mathcal{A} \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\ &= \bar{Q}^*_{M_1}(s,g,a) \quad (\text{since } r_{M_1}(g,a') \in \{r_{\text{MIN}}, r_{\text{MAX}}\}\ \forall a' \in \mathcal{A}). \end{aligned}$$

Similarly, $\bar{Q}^*_{MAX} \vee \bar{Q}^*_{M_1} = \bar{Q}^*_{MAX}$, $\bar{Q}^*_{MIN} \wedge \bar{Q}^*_{M_1} = \bar{Q}^*_{MIN}$, and $\bar{Q}^*_{MIN} \vee \bar{Q}^*_{M_1} = \bar{Q}^*_{M_1}$.

(vii): For all $(.)$ in $\mathcal{S} \times \mathcal{G} \times \mathcal{A}$,

$$\begin{aligned} (\bar{Q}^*_{M_1} \wedge \neg \bar{Q}^*_{M_1})(.) &= \min\{\bar{Q}^*_{M_1}(.), \neg \bar{Q}^*_{M_1}(.)\} \\ &= \begin{cases} \min\{\bar{Q}^*_{MIN}(.), \bar{Q}^*_{MAX}(.)\} & \text{if } |\bar{Q}^*(.) - \bar{Q}^*_{MIN}(.)| \leq |\bar{Q}^*(.) - \bar{Q}^*_{MAX}(.)| \\ \min\{\bar{Q}^*_{MAX}(.), \bar{Q}^*_{MIN}(.)\} & \text{otherwise,} \end{cases} \\ &= \bar{Q}^*_{MIN}(.). \end{aligned}$$

Similarly, $\bar{Q}^*_{M_1} \vee \neg \bar{Q}^*_{M_1} = \bar{Q}^*_{MAX}$.

□

Lemma 3. *Let $\bar{\mathcal{Q}}^*$ be the set of optimal extended $\bar{Q}$-value functions for tasks in $\mathcal{M}$. Then for all $M_1, M_2 \in \mathcal{M}$, we have (i) $\bar{Q}^*_{\neg M_1} = \neg \bar{Q}^*_{M_1}$, (ii) $\bar{Q}^*_{M_1 \vee M_2} = \bar{Q}^*_{M_1} \vee \bar{Q}^*_{M_2}$, and (iii) $\bar{Q}^*_{M_1 \wedge M_2} = \bar{Q}^*_{M_1} \wedge \bar{Q}^*_{M_2}$.*

Proof. Let $M_1, M_2 \in \mathcal{M}$. Then for all (s, g, a) in $\mathcal{S} \times \mathcal{G} \times \mathcal{A}$,

(i):

$$
\begin{aligned}
&\bar{Q}^*_{\neg_{M_1}}(s,g,a) \\
&= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } r_{\neg_{M_1}}(g,a') = r_{\text{MAX}}\ \forall a' \in \mathcal{A} \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\
&= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } r_{M_1}(g,a') = r_{\text{MIN}}\ \forall a' \in \mathcal{A} \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\
&= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } \bar{Q}^*_{M_1}(s,g,a) = \bar{Q}^*_{MIN}(s,g,a) \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\
&= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } |\bar{Q}^*_{M_1}(s,g,a) - \bar{Q}^*_{MIN}(s,g,a)| \leq |\bar{Q}^*_{M_1}(s,g,a) - \bar{Q}^*_{MAX}(s,g,a)| \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\
&= \neg\bar{Q}^*_{M_1}(s,g,a).
\end{aligned}
$$

(ii):

$$
\begin{aligned}
\bar{Q}^*_{M_1 \vee M_2}(s,g,a) &= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } r_{M_1 \vee M_2}(g,a') = r_{\text{MAX}}\ \forall a' \in \mathcal{A} \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\
&= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } \max\{r_{M_1}(g,a'), r_{M_2}(g,a')\} = r_{\text{MAX}}\ \forall a' \in \mathcal{A} \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\
&= \begin{cases} \bar{Q}^*_{MAX}(s,g,a), & \text{if } \max\{\bar{Q}^*_{M_1}(s,g,a), \bar{Q}^*_{M_2}(s,g,a)\} = \bar{Q}^*_{MAX}(s,g,a) \\ \bar{Q}^*_{MIN}(s,g,a), & \text{otherwise.} \end{cases} \\
&= \max\{\bar{Q}^*_{M_1}(s,g,a), \bar{Q}^*_{M_2}(s,g,a)\} \\
&= (\bar{Q}^*_{M_1} \vee \bar{Q}^*_{M_2})(s,g,a).
\end{aligned}
$$

(iii): Follows similarly to (ii).

□

Proposition 1. *Let $\bar{\mathcal{Q}}^*$ be the set of optimal $\bar{Q}$-value functions for tasks in $\mathcal{M}$. Let $\mathscr{A} : \mathcal{M} \to \bar{\mathcal{Q}}^*$ be any map from $\mathcal{M}$ to $\bar{\mathcal{Q}}^*$ such that $\mathscr{A}(M) = \bar{Q}^*_M$ for all M in $\mathcal{M}$. Then,*

- *(i) $\mathcal{M}$ and $\bar{\mathcal{Q}}^*$ respectively form a Boolean task algebra $(\mathcal{M}, \vee, \wedge, \neg, \mathcal{M}_{MAX}, \mathcal{M}_{MIN})$ and a Boolean EVF algebra $(\bar{\mathcal{Q}}^*, \vee, \wedge, \neg, \bar{Q}^*_{MAX}, \bar{Q}^*_{MIN})$,*
- *(ii) $\mathscr{A}$ is a homomorphism between $\mathcal{M}$ and $\bar{\mathcal{Q}}^*$.*

Proof. **(i):** Follows from Lemma 1 and 2.

(ii): Follows from Lemma 3.

□

1.3 Proofs for Theorem 1

Lemma 4. *Let $\bar{\mathcal{Q}}^*$ be the set of optimal $\bar{Q}$-value functions for tasks in $\mathcal{M}$. Denote ${}_M$ as the ϵ-optimal $\bar{Q}$-value function for a task $M \in \mathcal{M}$ such that*

$$|\bar{Q}^*_M(s,g,a) - {}_M\ (s,g,a)| \leq \epsilon \text{ for all } (s,g,a) \in \mathcal{S} \times \mathcal{G} \times \mathcal{A}.$$

Then for all M_1, M_2 in $\mathcal{M}$ and (s,g,a) in $\mathcal{S} \times \mathcal{G} \times \mathcal{A}$,

- *(i) $\left|[\bar{Q}^*_{M_1} \vee \bar{Q}^*_{M_2}](s,g,a) - [{}_{M_1} \vee {}_{M_2}](s,g,a)\right| \leq \epsilon$*
- *(ii) $\left|[\bar{Q}^*_{M_1} \wedge \bar{Q}^*_{M_2}](s,g,a) - [{}_{M_1} \wedge {}_{M_2}](s,g,a)\right| \leq \epsilon$*
- *(iii) $\left|\neg\bar{Q}^*_{M_1}(s,g,a) - \neg_{M_1}(s,g,a)\right| \leq \epsilon$*

Proof. **(i):**

$$\begin{aligned}
&\left|[\bar{Q}^*_{M_1} \vee \bar{Q}^*_{M_2}](s,g,a) - [_{M_1} \vee_{M_2}](s,g,a)\right| \\
&= \left| \max_{M \in \{M_1, M_2\}} \bar{Q}^*_M(s,g,a) - \max_{M \in \{M_1, M_2\}} {}_M(s,g,a) \right| \\
&\leq \max_{M \in \{M_1, M_2\}} \left|\bar{Q}^*_M(s,g,a) - {}_M\,(s,g,a)\right| \\
&\leq \epsilon.
\end{aligned}$$

(ii):

$$\begin{aligned}
&\left|[\bar{Q}^*_{M_1} \wedge \bar{Q}^*_{M_2}](s,g,a) - [_{M_1} \wedge_{M_2}](s,g,a)\right| \\
&= \left| \min_{M \in \{M_1, M_2\}} \bar{Q}^*_M(s,g,a) - \min_{M \in \{M_1, M_2\}} {}_M(s,g,a) \right| \\
&\leq \min_{M \in \{M_1, M_2\}} \left|\bar{Q}^*_M(s,g,a) - {}_M\,(s,g,a)\right| \\
&\leq \epsilon.
\end{aligned}$$

(iii):

$$\begin{aligned}
&\left|\neg \bar{Q}^*_{M_1}(s,g,a) - \neg_{M_1}(s,g,a)\right| \\
&= \begin{cases} |\bar{Q}^*_{MAX}(s,g,a) - \neg(s,g,a)|, & \text{if } \bar{Q}^*_{M_1} = \bar{Q}^*_{MIN}(s,g,a) \\ |\bar{Q}^*_{MIN}(s,g,a) - \neg(s,g,a)|, & \text{otherwise.} \end{cases} \\
&= \begin{cases} |\bar{Q}^*_{MAX}(s,g,a) - (s,g,a)|, & \text{if } \bar{Q}^*_{M_1} = \bar{Q}^*_{MIN}(s,g,a) \\ |\bar{Q}^*_{MIN}(s,g,a) - (s,g,a)|, & \text{otherwise.} \end{cases} \\
&\leq \epsilon.
\end{aligned}$$

□

Lemma 5. *Let $M \in \mathcal{M}$ be a task with binary specification T and optimal extended action-value function $\bar{Q}^*$. Given ϵ-approximations of the binary specifications $\tilde{\mathcal{T}}_n = \{\tilde{T}_1, ..., \tilde{T}_n\}$ and optimal $\bar{Q}$-functions $\tilde{\bar{\mathcal{Q}}}^*_n = \{\tilde{\bar{Q}}^*_1, ..., \tilde{\bar{Q}}^*_n\}$ for n tasks $\hat{\mathcal{M}} = \{M_1, ..., M_n\} \subseteq \mathcal{M}$, let*

$$T_{SOP} = \mathcal{B}_{EXP}(\tilde{\mathcal{T}}_n) \text{ and } \bar{Q}_{SOP} = \mathcal{B}_{EXP}(\tilde{\bar{\mathcal{Q}}}^*_n) \text{ where } \mathcal{B}_{EXP} = SOP(\tilde{\mathcal{T}}_n, \tilde{T}).$$

Define,

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} Q_{SOP} \text{ where } Q_{SOP} \coloneqq \max_{g \in \mathcal{G}} \bar{Q}_{SOP}(s,g,a).$$

Then,

$$\|\bar{Q}^* - \bar{Q}_{SOP}\|_\infty \leq (\mathbf{1}_{T \neq T_{SOP}}) r_\Delta + \epsilon,$$

*where **1** is the indicator function, $r_\Delta \coloneqq r_{MAX} - r_{MIN}$, and $\|f - h\|_\infty \coloneqq \max_{s,g,a} |f(s,g,a) - h(s,g,a)|$.*

Proof.

$$\begin{aligned}
|\bar{Q}^*(s,g,a) - \bar{Q}_{SOP}(s,g,a)| &= |\bar{Q}^*(s,g,a) - \bar{Q}^*_{SOP}(s,g,a) + \bar{Q}^*_{SOP}(s,g,a) - \bar{Q}_{SOP}(s,g,a)| \\
&\leq |\bar{Q}^*(s,g,a) - \bar{Q}^*_{SOP}(s,g,a)| + |\bar{Q}^*_{SOP}(s,g,a) - \bar{Q}_{SOP}(s,g,a)| \\
&\leq |\bar{Q}^*(s,g,a) - \bar{Q}^*_{SOP}(s,g,a)| + \epsilon. \qquad \text{(Using Lemma 4)}
\end{aligned}$$

If $T = T_{SOP}$, then $\bar{Q}^*(s,g,a) = \bar{Q}^*_{SOP}(s,g,a)$, and we are done. Let $T \neq T_{SOP}$. Without loss of generality, let $\bar{Q}^*(s,g,a) = \bar{Q}^*_{MAX}(s,g,a)$ and $\bar{Q}^*_{SOP}(s,g,a) = \bar{Q}^*_{MIN}(s,g,a)$. Then,

$$\begin{aligned}
|\bar{Q}^*(s,g,a) - \bar{Q}^*_{SOP}(s,g,a)| &\leq |\bar{Q}^*_{MAX}(s,g,a) - \bar{Q}^*_{MIN}(s,g,a)| \\
&\leq r_\Delta.
\end{aligned}$$

□

Lemma 6. *Let Q^* and $\bar{Q}^*$ be the optimal Q-value function and optimal extended Q-value function respectively for a deterministic task in $\mathcal{M}$. Then for all (s, a) in $\mathcal{S} \times \mathcal{A}$, we have*

$$Q^*(s,a) = \max_{g \in \mathcal{G}} \bar{Q}^*(s,g,a).$$

Proof. We first note that

$$\max_{g \in \mathcal{G}} \bar{r}(s,g,a) = \begin{cases} \max\{r_{\text{MIN}}, r(s,a)\}, & \text{if } s \in \mathcal{G} \\ \max_{g \in \mathcal{G}} r(s,a), & \text{otherwise.} \end{cases} = r(s,a). \tag{1}$$

Now define

$$\bar{Q}^*_{max}(s,a) \coloneqq \max_{g \in \mathcal{G}} \bar{Q}^*(s,g,a).$$

Then it follows that

$$\begin{aligned}
&\left[\bar{Q}^*_{max}\right](s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} \bar{Q}^*_{max}(s',a') \\
&= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} \left[\max_{g \in \mathcal{G}} \bar{Q}^*(s',g,a')\right] \\
&= r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{g \in \mathcal{G}} \left[\max_{a' \in \mathcal{A}} \bar{Q}^*(s',g,a')\right] \\
&= r(s,a) + \max_{g \in \mathcal{G}} \left[\gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} \bar{Q}^*(s',g,a')\right] \quad \text{(Since } p \text{ is deterministic)} \\
&= \max_{g \in \mathcal{G}} \bar{r}(s,g,a) + \max_{g \in \mathcal{G}} \left[\gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} \bar{Q}^*(s',g,a')\right] \quad \text{(Using Equation 1)} \\
&= \max_{g \in \mathcal{G}} \left[\bar{r}(s,g,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} \bar{Q}^*(s',g,a')\right], \\
&\text{since } \bar{r}(s,g,a) = r_0(s,a)\ \forall s \notin \mathcal{G} \text{ and } p(s,a,\omega) = 1 \text{ with } \bar{Q}^*(\omega,g,a') = 0\ \forall s \in \mathcal{G}. \\
&= \max_{g \in \mathcal{G}} \bar{Q}^*(s,g,a) \\
&= \bar{Q}^*_{max}(s,a).
\end{aligned}$$

Hence $\bar{Q}^*_{max}$ is a fixed point of the Bellman optimality operator.

If $s \in \mathcal{G}$, then

$$\bar{Q}^*_{max}(s,a) = \max_{g \in \mathcal{G}} Q^*(s,g,a) = \max_{g \in \mathcal{G}} \bar{r}(s,g,a) = r(s,a) = Q^*(s,a).$$

Since $\bar{Q}^*_{max} = Q^*$ holds in $\mathcal{G}$ and $\bar{Q}^*_{max}$ is a fixed point of the Bellman operator, then $\bar{Q}^*_{max} = Q^*$ holds everywhere.

□

Theorem 1. *Let $M \in \mathcal{M}$ be a task with binary specification T and optimal extended action-value function $\bar{Q}^*$. Given ϵ-approximations of the binary specifications $\tilde{\mathcal{T}}_n = \{\tilde{T}_1, ..., \tilde{T}_n\}$ and optimal $\bar{Q}$-functions $\tilde{\bar{\mathcal{Q}}}^*_n = \{\tilde{\bar{Q}}^*_1, ..., \tilde{\bar{Q}}^*_n\}$ for n tasks $\hat{\mathcal{M}} = \{M_1, ..., M_n\} \subseteq \mathcal{M}$, let*

$$T_{SOP} = \mathcal{B}_{EXP}(\tilde{\mathcal{T}}_n) \text{ and } \bar{Q}_{SOP} = \mathcal{B}_{EXP}(\tilde{\bar{\mathcal{Q}}}^*_n) \text{ where } \mathcal{B}_{EXP} = SOP(\tilde{\mathcal{T}}_n, \tilde{T}).$$

Define,

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} Q_{SOP} \text{ where } Q_{SOP} \coloneqq \max_{g \in \mathcal{G}} \bar{Q}_{SOP}(s,g,a).$$

Then,

(i) $\|Q^* - Q^\pi\|_\infty \leq \frac{2}{1-\gamma}((\boldsymbol{1}_{T \neq T_{SOP}} + \boldsymbol{1}_{r \notin \{r_g\}_{|\mathcal{G}|}})r_\Delta + \epsilon)$,

(ii) If the dynamics are deterministic,

$$\|Q^* - Q_{SOP}\|_\infty \leq (\boldsymbol{1}_{T \neq T_{SOP}})r_\Delta + \epsilon,$$

where $\boldsymbol{1}$ *is the indicator function,* $r_g(s,a) := \bar{r}(s,g,a)$*,* $r_\Delta := r_{MAX} - r_{MIN}$*, and* $\|f - h\|_\infty := \max_{s,g,a} |f(s,g,a) - h(s,g,a)|$.

Proof. **(i):** We first note that each g in $\mathcal{G}$ can be thought of as defining an MDP $M_g := (\mathcal{S}, \mathcal{A}, p, r_g, \gamma)$ with reward function $r_g(s,a) := \bar{r}(s,g,a)$, optimal policy $\pi_g^*(s) = \bar{\pi}^*(s,g)$ and optimal Q-value function $Q^{\pi_g^*}(s,a) = \bar{Q}^*(s,g,a)$. Then this proof follows similarly to that of Barreto et al. (2017) Theorem 2,

$$\begin{aligned}
&Q^*(s,a) - Q^\pi(s,a) \\
&\leq Q^*(s,a) - Q^{\pi_g^*}(s,a) + \frac{2}{1-\gamma}((\mathbf{1}_{T \neq T_{SOP}})r_\Delta + \epsilon) \quad \text{(Barreto et al. (2017) Theorem 1)} \\
&\leq \frac{2}{1-\gamma} \max_{s,a} |r(s,a) - r_g(s,a)| + \frac{2}{1-\gamma}((\mathbf{1}_{T \neq T_{SOP}})r_\Delta + \epsilon) \quad \text{(Barreto et al. (2017) Lemma 1)} \\
&\leq \frac{2}{1-\gamma}(\mathbf{1}_{r \neq r_g})r_\Delta + \frac{2}{1-\gamma}((\mathbf{1}_{T \neq T_{SOP}})r_\Delta + \epsilon) \\
&\quad \text{(Since rewards only differ in } \mathcal{G} \text{ where } r(s,a), r_g(s,a) \in \{r_{\text{MIN}}, r_{\text{MAX}}\} \text{ for } s \in \mathcal{G}) \\
&\leq \frac{2}{1-\gamma}((\mathbf{1}_{T \neq T_{SOP}} + \mathbf{1}_{r \neq r_g})r_\Delta + \epsilon).
\end{aligned}$$

Hence,

$$\begin{aligned}
\|Q^* - Q^\pi\|_\infty &\leq \frac{2}{1-\gamma}((\mathbf{1}_{T \neq T_{SOP}} + \min_g \mathbf{1}_{r \neq r_g})r_\Delta + \epsilon) \\
&\leq \frac{2}{1-\gamma}((\mathbf{1}_{T \neq T_{SOP}} + \mathbf{1}_{r \notin \{r_g\}_{|\mathcal{G}|}})r_\Delta + \epsilon) \\
&\quad \text{(Since } \min_g \mathbf{1}_{r \neq r_g} = 0 \text{ only when } r \in \{r_g\}_{|\mathcal{G}|}\text{)}.
\end{aligned}$$

(ii):

$$\begin{aligned}
|Q^*(s,a) - Q_{SOP}(s,a)| &= |\max_g \bar{Q}^*(s,g,a) - \max_g \bar{Q}_{SOP}(s,g,a)| \quad \text{(Lemma 6)} \\
&\leq \max_g |\bar{Q}^*(s,g,a) - \bar{Q}_{SOP}(s,g,a)| \\
&\leq (\mathbf{1}_{T \neq T_{SOP}})r_\Delta + \epsilon. \quad \text{(Lemma 5)}
\end{aligned}$$

□

1.4 Comparing the bounds of Theorem 1 with that of GPI in Barreto et al. (2018)

We first restate Proposition 1 (Barreto et al., 2018) here.

Proposition 2 ((Barreto et al., 2018))**.** *Let* $M \in \mathcal{M}$ *and let* $Q_i^{\pi_j^*}$ *be the action value function of an optimal policy of* $M_j \in \mathcal{M}$ *when executed in* $M_i \in \mathcal{M}$*. Given approximations* $\{\tilde{Q}_i^{\pi_1}, ..., \tilde{Q}_i^{\pi_n}\}$ *such that* $|Q_i^{\pi_j} - \tilde{Q}_i^{\pi_j}| \leq \epsilon$ *for all* $s, a \in \mathcal{S} \times \mathcal{A}$*, and* $j \in \{1, ..., n\}$*, let*

$$\pi(s) \in \arg\max_a \max_j \tilde{Q}_i^{\pi_j}(s,a).$$

then,

$$\|Q^* - Q^\pi\|_\infty \leq \frac{2}{1-\gamma}(\|r - r_i\|_\infty + \min_j \|r_i - r_j\|_\infty + \epsilon),$$

where Q^* *is the optimal value function of* M*,* Q^π *is the value function of* π *in* M*, and* $\|f - h\|_\infty := \max_{s,g,a} |f(s,g,a) - h(s,g,a)|$.

We can simplify the bound in Proposition 2 as follows:

$$\begin{aligned}
\|Q^* - Q^\pi\|_\infty &\leq \frac{2}{1-\gamma}(\|r - r_i\|_\infty + \min_j \|r_i - r_j\|_\infty + \epsilon) \\
&\leq \frac{2}{1-\gamma}((\mathbf{1}_{r \neq r_i})r_\Delta + \min_j \|r_i - r_j\|_\infty + \epsilon) \\
&\quad \text{(Since rewards only differ in } \mathcal{G} \text{ where } r(s,a), r_i(s,a) \in \{r_{\text{MIN}}, r_{\text{MAX}}\} \text{ for } s \in \mathcal{G}) \\
&\leq \frac{2}{1-\gamma}((\mathbf{1}_{r \neq r_i})r_\Delta + (\min_j \mathbf{1}_{r_i \neq r_j})r_\Delta + \epsilon) \\
&\leq \frac{2}{1-\gamma}((\mathbf{1}_{r \neq r_i})r_\Delta + (\mathbf{1}_{r_i \notin \{r_j\}_n})r_\Delta + \epsilon) \\
&\quad \text{(Since } \min_j \mathbf{1}_{r_i \neq r_j} = 0 \text{ only when } r_i \in \{r_j\}_n \text{)} \\
&\leq \frac{2}{1-\gamma}((\mathbf{1}_{r \neq r_i} + \mathbf{1}_{r_i \notin \{r_j\}_n})r_\Delta + \epsilon).
\end{aligned}$$

where $\mathbf{1}$ is the indicator function, and $r_\Delta \coloneqq r_{\text{MAX}} - r_{\text{MIN}}$. We can see that this bound is similar to that of Theorem 1(i) but weaker. This because:

(i) The first term of this bound requires that the current task be identical to the task being approximated—$\mathbf{1}_{r \neq r_i}$—while the first term of Theorem 1(i) only requires the current task to be expressible as a Boolean composition of past tasks—$\mathbf{1}_{T \neq T_{SOP}}$.

(ii) The second term of this bound requires that the task being approximated is one of the past tasks—$\mathbf{1}_{r_i \notin \{r_j\}_n}$—while the second term of Theorem 1(i) only requires the current task to have a single desirable goal—$\mathbf{1}_{r \notin \{r_g\}_{|\mathcal{G}|}}$.

(iii) Barreto et al. (2018) assumes that the reward function of the current task is well approximated by a linear function over a fixed set of rewards. Hence while a new task may be expressed as the Boolean composition of past tasks—$T = T_{SOP}$—, its rewards may not be expressible as a linear combination of a fixed set of rewards—$r \neq r_i$ where $r_i \coloneqq [r_0, ..., r_n] * w$.

This suggests that we can can think of the SOP composition approach as an efficient way of doing GPI, one which leads to tight performance bounds on the transferred policy (Theorem 1(ii)).

1.5 Proofs for Theorem 2

Theorem 2. *Let $\mathcal{D}$ be an unknown non-stationary distribution over a set of tasks $\mathcal{M}(\mathcal{S}, \mathcal{A}, p, \gamma, r_0)$, and let $\mathscr{A} : \mathcal{M} \to \bar{\mathcal{Q}}^*$ be any map from $\mathcal{M}$ to $\bar{\mathcal{Q}}^*$ such that $\mathscr{A}(M) = \bar{Q}^*_M$ for all M in $\mathcal{M}$. Let*

$$\tilde{\mathcal{T}}_{t+1}, \tilde{\bar{\mathcal{Q}}}^*_{t+1} = SOPGOL(\mathscr{A}, M_t, \tilde{\mathcal{T}}_t, \tilde{\bar{\mathcal{Q}}}^*_t) \text{ where } M_t \sim \mathcal{D}(t) \text{ and } \tilde{\mathcal{T}}_0 = \tilde{\bar{\mathcal{Q}}}^*_0 = \varnothing \; \forall t \in \mathbb{N}.$$

Then,

$$\lceil \log |\mathcal{G}| \rceil \leq \lim_{t \to \infty} |\tilde{\mathcal{T}}_t| = \lim_{t \to \infty} |\tilde{\bar{\mathcal{Q}}}^*_t| \leq |\mathcal{G}|.$$

Proof. Let $\tilde{T}_t$ be the approximate binary specification of task M_t learned by SOPGOL. We first note that SOPGOL returns $\tilde{\mathcal{T}}_t \cup \{\tilde{T}_t\}$ only if $\tilde{T}_t$ is not in the span of $\tilde{\mathcal{T}}_t$. That is,

$$\tilde{\mathcal{T}}_{t+1} = \tilde{\mathcal{T}}_t \cup \{\tilde{T}_t\} \textit{ iff } \tilde{T}_t \neq \mathcal{B}_{EXP}(\tilde{\mathcal{T}}_t) \text{ where } \mathcal{B}_{EXP} = SOP(\tilde{\mathcal{T}}_t, \tilde{T}_t).$$

Hence, it is sufficient to show that the number, N, of linearly independent binary vectors, $\tilde{T} \in \{0,1\}^{|\mathcal{G}|}$, that span the Boolean vector space (Subrahmanyam, 1964), $GF(2)^{|\mathcal{G}|}$,[1] is bounded by

$$\lceil \log |\mathcal{G}| \rceil \leq N \leq |\mathcal{G}|.$$

This follows from the fact that $\lceil \log |\mathcal{G}| \rceil$ is the size of a minimal basis of $GF(2)^{|\mathcal{G}|}$ (as can easily be seen with a Boolean table), and $|\mathcal{G}|$ is its dimensionality.

□

[1] GF(2) is the Galois field with two elements, $(\{0,1\}, +, .)$, where $+ \coloneqq XOR$ and $. \coloneqq AND$.

2 SUM OF PRODUCTS WITH GOAL ORIENTED LEARNING

Algorithm 1: SOPGOL

Input : off-policy RL algorithm $\mathscr{A}$, `/* e.g DQN */`
task MDP M,
set of ϵ-optimal task binary specifications $\tilde{\mathcal{T}}$,
set of ϵ-optimal $\bar{Q}$-value functions $\tilde{\bar{\mathcal{Q}}}$.

Initialise $\tilde{T} : \mathcal{G} \to \{0, 1\}$

Initialise $\tilde{\bar{Q}} : \mathcal{S} \times \mathcal{G} \times \mathcal{A} \to \mathbb{R}$ according to $\mathscr{A}$

Initialise goal buffer $\tilde{\mathcal{G}}$ with terminal states observed from a random policy

while $\tilde{\bar{Q}}$ *is not converged* **do**
  Initialise state s from M
  $\mathcal{B}_{EXP} \leftarrow SOP(\tilde{\mathcal{T}}, \tilde{T})$
  $T_{SOP}, \bar{Q}_{SOP} \leftarrow \mathcal{B}_{EXP}(\tilde{\mathcal{T}}), \mathcal{B}_{EXP}(\tilde{\bar{\mathcal{Q}}}^*)$
  $\bar{Q} \leftarrow \bar{Q}_{SOP}$ **if** $\tilde{T} = T_{SOP}$ **else** $\tilde{\bar{Q}} \vee \bar{Q}_{SOP}$
  $g \leftarrow \arg\max_{g' \in \tilde{\mathcal{G}}} \left(\max_{a \in \mathcal{A}} \bar{Q}(s, g', a) \right)$
  while s *is not terminal* **do**
    Select action a using the behaviour policy from $\mathscr{A}$: $a \leftarrow \bar{\pi}(s, g)$ `/* e.g ϵ-greedy */`
    Take action a, observe reward r and next state s' in M
    if $\tilde{T} \neq T_{SOP}$ **then**
      foreach $g' \in \tilde{\mathcal{G}}$ **do**
        $\bar{r} \leftarrow r_{\text{MIN}}$ **if** $g' \neq s \in \tilde{\mathcal{G}}$ **else** r
        Update $\tilde{\bar{Q}}$ with $(s, g', a, \bar{r}, s')$ according to $\mathscr{A}$
      end
    if s *is terminal* **then**
      $\tilde{T}(s) \leftarrow \mathbf{1}_{r=r_{\text{MAX}}}$
      $\tilde{\mathcal{G}} \leftarrow \tilde{\mathcal{G}} \cup \{s\}$
    else
      $s \leftarrow s'$
    end
  end
end
$\mathcal{B}_{EXP} \leftarrow SOP(\tilde{\mathcal{T}}, \tilde{T})$
$\tilde{\mathcal{T}}, \tilde{\bar{\mathcal{Q}}} \leftarrow (\tilde{\mathcal{T}}, \tilde{\bar{\mathcal{Q}}})$ **if** $\tilde{T} = \mathcal{B}_{EXP}(\tilde{\mathcal{T}})$ **else** $(\tilde{\mathcal{T}} \cup \{\tilde{T}\}, \tilde{\bar{\mathcal{Q}}} \cup \{\tilde{\bar{Q}}\})$
return $\tilde{\mathcal{T}}, \tilde{\bar{\mathcal{Q}}}$

3 FUNCTION APPROXIMATION EXPERIMENT DETAILS

3.1 ENVIRONMENT

The PickUpObj environment is fully observable, where each state observation is a $56 * 56 * 3$ RGB image (Figure 1). The agent has 7 actions it can take in this environment corresponding to: 1 - rotate left, 2 - rotate right, 3 - move one step forward if there is no wall or object in front, 4 - pickup object if there is an object in front and no object has been picked, 5 - drop the object in front if an object has been picked and there is no wall or object in front, 6 - open the door in front if there is a closed-door in front, and 7 - close the door in front if there is an opened door in front.

For each task, each episode starts with 1 desirable object and 4 other randomly chosen objects placed randomly in the environment. The agent is also placed at a random position with a random orientation at the start of each episode. The agent receives a reward of -0.1 at every timestep, and a reward of 2 when it picks up a desirable object. The environment transitions to a terminal state once the agent picks up any object and the agent observes the picked object. There are 15 types of objects (illustrated in Table 1) resulting in 15 possible goal states. Hence, the dimension of the state space is $|\mathcal{S}| = 56 * 56 * 3$, the goal space is $|\mathcal{G}| = 15$, and the action space is $|\mathcal{A}| = 7$.

3.2 NETWORK ARCHITECTURE AND HYPERPARAMETERS

In our function approximation experiments, we represent each extended value function $\tilde{\tilde{Q}}^*$ with a list of $|\mathcal{G}|$ DQNs, such that the value function for each goal $\tilde{Q}^*_g(s,a) \coloneqq \tilde{\tilde{Q}}^*(s,g,a)$ is approximated with a separate DQN. The DQNs used have the following architecture, with the CNN part being identical to that used by Mnih et al. (2015):

1. Three convolutional layers:
   (a) Layer 1 has 3 input channels, 32 output channels, a kernel size of 8 and a stride of 4.
   (b) Layer 2 has 32 input channels, 64 output channels, a kernel size of 4 and a stride of 2.
   (c) Layer 3 has 64 input channels, 64 output channels, a kernel size of 3 and a stride of 1.
2. Two fully-connected linear layers:
   (a) Layer 1 has input size 3136 and output size 512 and uses a ReLU activation function.
   (b) Layer 2 has input size 512 and output size 7 with no activation function.

We used the ADAM optimiser with batch size 256 and a learning rate of 10^{-3}. We started training after 1000 steps of random exploration and updated the target Q-network every 1000 steps. Finally, we used ϵ-greedy exploration, annealing ϵ from 0.5 to 0.05 over 100000 timesteps.

Finally, we used the same DQN architecture and training hyperparameters for the baseline in all experiments.

REFERENCES

A. Barreto, W. Dabney, R. Munos, J. Hunt, T. Schaul, H. van Hasselt, and D. Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4055–4065, 2017.

Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International Conference on Machine Learning*, pp. 501–510. PMLR, 2018.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

NV Subrahmanyam. Boolean vector spaces. *Mathematische Zeitschrift*, 83(5):422–433, 1964.