A Proofs

In this section, we prove the theorems presented in the paper through a series of lemmas. The proofs here are an adaptation of proofs of similar results for zero-sum concurrent games with a single discount factor [34].

A.1 Definitions

We first introduce some notation and definitions. Recall that we defined $S_1 = \{(s, \sigma) \mid \sigma \notin F_\sigma\}$, $S_2 = \{(s, \sigma) \mid \sigma \in F_\sigma\}$ and $\bar{S} = S_1 \cup S_2$. $\mathcal{V} = \{V : S_1 \to \mathbb{R}\}$ denotes the set of value functions over S_1 and $\bar{\mathcal{V}} = \{V : \bar{S} \to \mathbb{R}\}$ denotes the set of value functions over $\bar{S}$. We use $\|\cdot\|$ to denote the ℓ_∞-norm. For any $V \in \mathcal{V}$, $(s, \sigma) \in S_2$ and any $\sigma' \in \Sigma$, define

$$[\![V]\!]_{\sigma'}(s, \sigma) = \sum_{s' \in S} T_\sigma(s' \mid s) V(s', \sigma').$$

Note that for any $(s, \sigma) \in S_2$, we have $[\![V]\!](s, \sigma) = \min_{\sigma' \in \Sigma} [\![V]\!]_{\sigma'}(s, \sigma)$. Similarly, for any $V \in \mathcal{V}$, $(s, \sigma) \in S_1$ and $a \in A$, let us define

$$\mathcal{F}_a(V)(s, \sigma) = \bar{R}((s, \sigma), a) + \gamma \cdot \sum_{s' \in S} P(s' \mid s, a) [\![V]\!](s', \sigma)$$

with $\mathcal{F}(V)(s, \sigma) = \max_{a \in A} \mathcal{F}_a(V)(s, \sigma)$. Given any policy $\pi_1 : \bar{S} \to A_1$ for agent 1 in $\mathcal{G}$, we define the resulting MDP $\mathcal{G}(\pi_1) = (\bar{S}, A_2, P_{\pi_1}, R_{\pi_1}, \gamma)$ with states $\bar{S}$ and actions $\mathcal{A}_2 = \Sigma$ as follows. The transition probability function is given by

$$P_{\pi_1}((s', \sigma') \mid (s, \sigma), a_2) = \begin{cases} \bar{P}((s', \sigma') \mid (s, \sigma), \pi_1(s, \sigma), a_2) & \text{if } (s, \sigma) \in S_1 \\ \sum_{s'' \in S} T_\sigma(s'' \mid s) \bar{P}((s', \sigma') \mid (s'', a_2), \pi_1(s'', a_2), a_2) & \text{if } (s, \sigma) \in S_2 \end{cases}$$

and the reward function is given by

$$R_{\pi_1}((s, \sigma), a_2) = \begin{cases} -\bar{R}((s, \sigma), \pi_1(s, \sigma)) & \text{if } (s, \sigma) \in S_1 \\ -\sum_{s' \in S} T_\sigma(s' \mid s) \bar{R}((s', a_2), \pi_1(s', a_2)) & \text{if } (s, \sigma) \in S_2. \end{cases}$$

Intuitively, the MDP $\mathcal{G}(\pi_1)$ merges every step of $\mathcal{G}$ in which a change of subtask occurs with the subsequent step in the environment, while using π_1 to choose actions for agent 1. For any $\bar{s} \in \bar{S}$, let $\mathcal{D}_{\bar{s}}^{\mathcal{G}(\pi_1)}(\pi_2)$ denote the distribution over infinite trajectories generated by π_2 starting at $\bar{s}$ in $\mathcal{G}(\pi_1)$. Then we define the value function for the MDP $\mathcal{G}(\pi_1)$ using

$$V_{\mathcal{G}(\pi_1)}^{\pi_2}(\bar{s}) = \mathbb{E}_{\rho \sim \mathcal{D}_{\bar{s}}^{\mathcal{G}(\pi_1)}(\pi_2)} \Big[\sum_{t=0}^{\infty} \gamma^t R_{\pi_1}(\bar{s}_t, a_t) \Big]$$

for all $\pi_2 : \bar{S} \to A_2$ and $\bar{s} \in \bar{S}$.

A.2 Necessary Lemmas

We need a few intermediate results in order to prove the main theorems. We begin by analyzing the operators $[\![\cdot]\!] : \mathcal{V} \to \bar{\mathcal{V}}$ and $\mathcal{F} : \mathcal{V} \to \mathcal{V}$ defined in Section 3.1.

Lemma A.1. *For any $V_1, V_2 \in \mathcal{V}$, we have $\|[\![V_1]\!] - [\![V_2]\!]\| = \|V_1 - V_2\|$.*

Proof. For any $(s, \sigma) \in S_1$, we have $|[\![V_1]\!](s, \sigma) - [\![V_2]\!](s, \sigma)| = |V_1(s, \sigma) - V_2(s, \sigma)|$. For any $(s, \sigma) \in S_2$ and $\sigma' \in \Sigma$ we have

$$\begin{aligned} |[\![V_1]\!]_{\sigma'}(s, \sigma) - [\![V_2]\!]_{\sigma'}(s, \sigma)| &= \|\sum_{s' \in S} T_\sigma(s' \mid s) V_1(s', \sigma') - \sum_{s' \in S} T_\sigma(s' \mid s) V_2(s', \sigma')\| \\ &\leq \sum_{s' \in S} T_\sigma(s' \mid s) |V_1(s', \sigma') - V_2(s', \sigma')| \\ &\leq \|V_1 - V_2\|. \end{aligned}$$

Now we have $|[\![V_1]\!](s, \sigma) - [\![V_2]\!](s, \sigma)| = |\min_{\sigma'} [\![V_1]\!]_{\sigma'}(s, \sigma) - \min_{\sigma'} [\![V_2]\!]_{\sigma'}(s, \sigma)| \leq \|V_1 - V_2\|$ which concludes the proof. □

Now we are ready to show that $\mathcal{F}$ is a contraction.

Lemma A.2. *$\mathcal{F} : \mathcal{V} \to \mathcal{V}$ is a contraction mapping w.r.t the norm $\|\cdot\|$.*

Proof. Let $V_1, V_2 \in \mathcal{V}$. Then for any $(s, \sigma) \in S_1$ and $a \in A$,

$$\begin{aligned}
|\mathcal{F}_a(V_1)(s,\sigma) - \mathcal{F}_a(V_2)(s,\sigma)| &= |\gamma \sum_{s' \in S} P(s' \mid s, a)[\![V_1]\!](s', \sigma) - \gamma \sum_{s' \in S} P(s' \mid s, a)[\![V_2]\!](s', \sigma)| \\
&\leq \gamma \sum_{s' \in S} P(s' \mid s, a)|[\![V_1]\!](s', \sigma) - [\![V_2]\!](s', \sigma)| \\
&\leq \gamma \|V_1 - V_2\|
\end{aligned}$$

where the last inequality followed from Lemma A.1. Therefore, for any $(s, \sigma) \in S_1$, we have $|\mathcal{F}(V_1)(s,\sigma) - \mathcal{F}(V_2)(s,\sigma)| = |\max_a \mathcal{F}_a(V_1)(s,\sigma) - \max_a \mathcal{F}_a(V_2)(s,\sigma)| \leq \gamma \|V_1 - V_2\|$ showing that $\mathcal{F}$ is a contraction. □

Now we connect the value function of the game $\mathcal{G}$ with that of the MDP $\mathcal{G}(\pi_1)$.

Lemma A.3. *For any $\pi_1 : \bar{S} \to A_1$, $\pi_2 : \bar{S} \to A_2$ and $\bar{s} \in \bar{S}$, $V^{\pi_1,\pi_2}(\bar{s}) = -V^{\pi_2}_{\mathcal{G}(\pi_1)}(\bar{s})$.*

Proof. Given an infinite trajectory $\bar{\rho} = \bar{s}_0 a^1_0 a^2_0 \bar{s}_1 a^1_1 a^2_1 \ldots$ in $\mathcal{G}$ we define a corresponding trajectory $\bar{\rho}_2 = \bar{s}_{i_0} a^2_{i_0} \bar{s}_{i_1} a^2_{i_1}$ in $\mathcal{G}(\pi_1)$ as a subsequence where $i_0 = 0$ and $i_{t+1} = i_t + 1$ if $\bar{s}_{i_t} \in S_1$ and $i_{t+1} = i_t + 2$ if $\bar{s}_{i_t} \in S_2$. Then for any $\bar{s} \in \bar{S}$ we have

$$\begin{aligned}
V^{\pi_1,\pi_2}(\bar{s}) &= \mathbb{E}_{\bar{\rho} \sim \mathcal{D}^{\mathcal{G}}_{\bar{s}}(\pi_1,\pi_2)}\Big[\sum_{t=0}^{\infty}\big(\prod_{k=0}^{t-1}\bar{\gamma}(\bar{s}_k)\big)\bar{R}(\bar{s}_t, a^1_t)\Big] \\
&\overset{(1)}{=} \mathbb{E}_{\bar{\rho} \sim \mathcal{D}^{\mathcal{G}}_{\bar{s}}(\pi_1,\pi_2)}\Big[\sum_{t=0}^{\infty}\gamma^t R_{\pi_1}(\bar{s}_{i_t}, a^2_{i_t})\Big] \\
&\overset{(2)}{=} -\mathbb{E}_{\rho \sim \mathcal{D}^{\mathcal{G}(\pi_1)}_{\bar{s}}(\pi_2)}\Big[\sum_{t=0}^{\infty}\gamma^t R_{\pi_1}(\bar{s}_t, a_t)\Big] \\
&= -V^{\pi_2}_{\mathcal{G}(\pi_1)}(\bar{s})
\end{aligned}$$

where (1) followed from the definitions of $\bar{\gamma}$ and R_{π_1} and the fact that $\bar{R}(\bar{s}_t, a^1_t) = 0$ if $\bar{s}_t \in S_2$, and (2) followed from the fact that sampling a trajectory ρ by first sampling $\bar{\rho}$ from $\mathcal{D}^{\mathcal{G}}_{\bar{s}}(\pi_1, \pi_2)$ and then constructing the subsequence $\bar{\rho}_2$ is the same as sampling an infinite trajectory ρ from $\mathcal{D}^{\mathcal{G}(\pi_1)}_{\bar{s}}(\pi_2)$[6]. □

Lemma A.2 shows that for any $V \in \mathcal{V}$ we have

$$\lim_{n \to \infty} \mathcal{F}^n(V) = V_{\lim}$$

where $V_{\lim} \in \mathcal{V}$ is the unique fixed point of $\mathcal{F}$. Now we define two policies π^*_1 and π^*_2 for agents 1 and 2 respectively, as follows. For $(s, \sigma) \in S_1$ we have

$$\pi^*_1(s, \sigma) \in \arg\max_{a \in A} \mathcal{F}_a(V_{\lim})(s, \sigma) \tag{5}$$

and for $(s, \sigma) \in S_2$, we have

$$\pi^*_2(s, \sigma) \in \arg\min_{\sigma'} [\![V_{\lim}]\!]_{\sigma'}(s, \sigma). \tag{6}$$

Note that the actions taken by π^*_1 in S_2 and π^*_2 in S_1 can be arbitrary since they do not affect the transitions of the game $\mathcal{G}$. Now we show that for any $\bar{s} \in \bar{S}$, π^*_1 maximizes $V^{\pi_1,\pi^*_2}(\bar{s})$ and π^*_2 minimizes $V^{\pi^*_1,\pi_2}(\bar{s})$.

Lemma A.4. *For any $\bar{s} \in \bar{S}$, $V^{\pi^*_1,\pi^*_2}(\bar{s}) = \min_{\pi_2} V^{\pi^*_1,\pi_2}(\bar{s}) = [\![V_{\lim}]\!](\bar{s})$.*

[6]This can be shown formally by analyzing the probabilities assigned by the two distributions on cylinder sets.

Proof. Let $\mathcal{G}(\pi_1^*) = (\bar{S}, \mathcal{A}_2, P_{\pi_1^*}, R_{\pi_1^*}, \gamma)$. For any $(s, \sigma) \in S_2$, we have

$$\begin{aligned}
[\![V_{\lim}]\!](s,\sigma) &= \min_{\sigma' \in \Sigma} \sum_{s' \in S} T_\sigma(s' \mid s)[\![V_{\lim}]\!](s', \sigma') \\
&\overset{(3)}{=} \min_{\sigma' \in \Sigma} \sum_{s' \in S} T_\sigma(s' \mid s)\Big(\bar{R}((s',\sigma'),a) + \gamma \cdot \sum_{s'' \in S} P(s' \mid s, a)[\![V_{\lim}]\!](s'', \sigma')\Big)\Big|_{a=\pi_1^*(s',\sigma')} \\
&\overset{(4)}{=} \min_{a_2 \in A_2} \Big(- R_{\pi_1^*}((s,\sigma), a_2) + \gamma \sum_{\bar{s} \in \bar{S}} P_{\pi_1^*}(\bar{s} \mid (s,\sigma), a_2)[\![V_{\lim}]\!](\bar{s})\Big)
\end{aligned}$$

where (3) followed from the definitions of $V_{\lim}$ and π_1^* and (4) followed from the definitions of $R_{\pi_1^*}$ and $P_{\pi_1^*}$. Since $-[\![V_{\lim}]\!]$ satisfies the Bellman equations for the MDP $\mathcal{G}(\pi_1^*)$, the optimal value function for $\mathcal{G}(\pi_1^*)$ is given by $V^*_{\mathcal{G}(\pi_1^*)} = -[\![V_{\lim}]\!]$. Now, from the definition of π_2^* we can conclude that π_2^* is an optimal policy for $\mathcal{G}(\pi_1^*)$. Therefore, Lemma A.3 implies that for any $\bar{s} \in \bar{S}$,

$$\min_{\pi_2} V^{\pi_1^*, \pi_2}(\bar{s}) = \min_{\pi_2} -V^{\pi_2}_{\mathcal{G}(\pi_1^*)}(\bar{s}) = -V^{\pi_2^*}_{\mathcal{G}(\pi_1^*)}(\bar{s}) = V^{\pi_1^*, \pi_2^*}(\bar{s}).$$

Hence, we have proved the desired result. □

The following lemma can be shown using a similar argument and the proof is omitted.

Lemma A.5. *For any $\bar{s} \in \bar{S}$, $V^{\pi_1^*, \pi_2^*}(\bar{s}) = \max_{\pi_1} V^{\pi_1, \pi_2^*}(\bar{s}) = [\![V_{\lim}]\!](\bar{s})$.*

The following lemma shows that it does not matter which agent picks its policy first.

Lemma A.6. *For any policies π_1^* and π_2^* satisfying Equations 5 and 6 respectively, for all $\bar{s} \in \bar{S}$,*

$$V^{\pi_1^*, \pi_2^*}(\bar{s}) = \min_{\pi_2} \max_{\pi_1} V^{\pi_1, \pi_2}(\bar{s}) = \max_{\pi_1} \min_{\pi_2} V^{\pi_1, \pi_2}(\bar{s}) = V^*(\bar{s}).$$

Proof. We have, for any $\bar{s} \in \bar{S}$,

$$\begin{aligned}
V^*(\bar{s}) &= \max_{\pi_1} \min_{\pi_2} V^{\pi_1, \pi_2}(\bar{s}) \\
&\geq \min_{\pi_2} V^{\pi_1^*, \pi_2}(\bar{s}) \\
&\overset{(1)}{=} V^{\pi_1^*, \pi_2^*}(\bar{s}) \\
&\overset{(2)}{=} \max_{\pi_1} V^{\pi_1, \pi_2^*}(\bar{s}) \\
&\geq \min_{\pi_2} \max_{\pi_1} V^{\pi_1, \pi_2}(\bar{s}) \\
&\geq \max_{\pi_1} \min_{\pi_2} V^{\pi_1, \pi_2}(\bar{s}) \\
&= V^*(\bar{s})
\end{aligned}$$

where the (1) followed from Lemma A.4 and (2) followed from Lemma A.5. □

A.3 Proof of Theorem 3.1

Let $\Pi(\pi_1) = \{\pi_\sigma \mid \sigma \in \Sigma\}$ be the set of subtask policies defined by π_1. Let $\tau = \sigma_0 \sigma_1 \ldots$ be a task. Then we define a history-dependent policy π_2^τ in $\mathcal{G}(\pi_1)$ which maintains an index i denoting the current subtask and picks σ_{i+1} upon reaching any state in S_2 while simultaneously updating the

index to $i+1$. Then we have

$$\begin{aligned}
J(\Pi(\pi_1)) &= \inf_{\tau\in\mathcal{T}} \mathbb{E}_{\rho\sim\mathcal{D}_\tau^\Pi}\Big[\sum_{t=0}^{\infty}\gamma^t R_{\tau[i_t]}(s_t,\pi_{\tau[i_t]}(s_t))\Big] \\
&\overset{(1)}{=} \inf_{\tau\in\mathcal{T}} \mathbb{E}_{\bar{s}\sim\bar{\eta}}\Big[\mathbb{E}_{\rho\sim\mathcal{D}_{\bar{s}}^{\mathcal{G}(\pi_1)}(\pi_2^\tau)}\big[-\sum_{t=0}^{\infty}\gamma^t R_{\pi_1}(\bar{s}_t,a_t)\big]\Big] \\
&= \inf_{\tau\in\mathcal{T}} \mathbb{E}_{\bar{s}\sim\bar{\eta}}[-V_{\mathcal{G}(\pi_1)}^{\pi_2^\tau}(\bar{s})] \\
&\geq \mathbb{E}_{\bar{s}\sim\bar{\eta}}[-\sup_{\tau\in\mathcal{T}} V_{\mathcal{G}(\pi_1)}^{\pi_2^\tau}(\bar{s})] \\
&\overset{(2)}{\geq} \mathbb{E}_{\bar{s}\sim\bar{\eta}}[-\max_{\pi_2} V_{\mathcal{G}(\pi_1)}^{\pi_2}(\bar{s})] \\
&\overset{(3)}{=} \mathbb{E}_{\bar{s}\sim\bar{\eta}}[\min_{\pi_2} V^{\pi_1,\pi_2}(\bar{s})] \\
&= J_{\mathcal{G}}(\pi_1)
\end{aligned}$$

where (1) followed from the definitions of π_2^τ and $\mathcal{G}(\pi_1)$, (2) followed from the fact that there is an optimal stationary policy maximizing $V_{\mathcal{G}(\pi_1)}^{\pi_2}(\bar{s})$ and (3) followed from Lemma A.4.

A.4 Proof of Theorem 3.2

Since $V^* = [\![V_{\lim}]\!]$, for all $(s,\sigma)\in\bar{S}$ we have $\pi_1^*(s,\sigma)\in\arg\max_{a\in A}\mathcal{F}_a(V_{\lim})(s,\sigma)$. Now for any π_2^* satisfying Equation 6, we can conclude from Lemma A.6 that, for any $\bar{s}\in\bar{S}$,

$$\begin{aligned}
J_{\mathcal{G}}(\pi_1^*) &= \mathbb{E}_{\bar{s}\sim\bar{\eta}}[\min_{\pi_2} V^{\pi_1^*,\pi_2}(\bar{s})] \\
&= \mathbb{E}_{\bar{s}\sim\bar{\eta}}[V^{\pi_1^*,\pi_2^*}(\bar{s})] \\
&= \mathbb{E}_{\bar{s}\sim\bar{\eta}}[\max_{\pi_1}\min_{\pi_2} V^{\pi_1,\pi_2}(\bar{s})] \\
&\geq \max_{\pi_1}\mathbb{E}_{\bar{s}\sim\bar{\eta}}[\min_{\pi_2} V^{\pi_1,\pi_2}(\bar{s})] \\
&= \max_{\pi_1} J_{\mathcal{G}}(\pi_1)
\end{aligned}$$

which shows that π_1^* maximizes $J_{\mathcal{G}}(\pi_1)$. □

A.5 Proof of Theorem 3.3

From Lemma A.2, we can conclude that $\mathcal{F}$ is a contraction over $\mathcal{V}$ w.r.t. the ℓ_∞-norm. Lemmas A.6 and A.4 gives us that $V^*\downarrow_{S_1} = V_{\lim}$. Now the definition of $V_{\lim}$ implies that $\lim_{n\to\infty}\mathcal{F}^n(V) = V^*\downarrow_{S_1}$ for all $V\in\mathcal{V}$. □

A.6 Proof of Theorems 3.4 and 3.5

This proof is similar to the proof of convergence of asynchronous value iteration for MDPs presented in [7]. It is easy to see that, for any $V\in\mathcal{V}$ and $\sigma\in\Sigma$, the operators $[\![\cdot]\!]$, $\mathcal{F}$, $\mathcal{F}_{\texttt{async}}$, and $\mathcal{F}_{\sigma,V}$ are monotonic. Recall that, for any $V\in\mathcal{V}$ and $\sigma\in\Sigma$, we defined the corresponding $V_\sigma\in\mathcal{V}_\sigma$ using $V_\sigma(s) = [\![V]\!](s,\sigma)$ if $s\in S$ and $V_\sigma(\bot) = 0$. Also, we have $\mathcal{F}(V)(s,\sigma) = \mathcal{F}_{\sigma,V}(V_\sigma)(s) = \mathcal{F}_1(V)(s,\sigma)$ for all $(s,\sigma)\in S_1$.

Now let $V\in\mathcal{V}$ be a value function such that $\mathcal{F}(V)\leq V$. Then we have $\mathcal{F}_{\sigma,V}(V_\sigma)\leq V_\sigma$ for all $\sigma\in\Sigma$. Therefore, using monotonicity of $\mathcal{F}_{\sigma,V}$, we get that $\mathcal{F}_{\sigma,V}^m(V_\sigma)\leq\mathcal{F}_{\sigma,V}^{m-1}(V_\sigma)\leq V_\sigma$ for all $m>0$ which implies $\mathcal{F}_m(V)\leq\mathcal{F}_{m-1}(V)\leq V$. Hence, for any $(s,\sigma)\in S_1$,

$$\begin{aligned}
\mathcal{F}_{\texttt{async}}(V)(s,\sigma) = \mathcal{W}_\sigma(V)(s) &= \lim_{m\to\infty}\mathcal{F}_{\sigma,V}^m(V_\sigma)(s) \\
&\leq \mathcal{F}_{\sigma,V}(V_\sigma)(s) = \mathcal{F}(V)(s,\sigma).
\end{aligned}$$

Furthermore, letting $V^m = \mathcal{F}_m(V)$ we get that $[\![V^m]\!]\leq[\![V]\!]$ and hence $V_\sigma^m\leq\mathcal{F}_{\sigma,V}^m(V_\sigma)$ for all $\sigma\in\Sigma$. Also, for $(s,\sigma)\in S_1$, $\mathcal{F}(V^m)(s,\sigma) = \mathcal{F}_{\sigma,V^m}(V_\sigma^m)(s)\leq\mathcal{F}_{\sigma,V}(V_\sigma^m)(s)\leq\mathcal{F}_{\sigma,V}^{m+1}(V_\sigma)(s) =$

$V^{m+1}(s,\sigma)$. Therefore, using continuity of $\mathcal{F}$, we have $\mathcal{F}(\mathcal{F}_{\texttt{async}}(V)) = \mathcal{F}(\lim_{m\to\infty} V^m) = \lim_{m\to\infty} \mathcal{F}(V^m) \leq \lim_{m\to\infty} V^{m+1} = \mathcal{F}_{\texttt{async}}(V)$. Now we can show by induction on n that, for any $V \in \mathcal{V}$ with $\mathcal{F}(V) \leq V$ and $n \geq 1$, we have $\mathcal{F}(\mathcal{F}^n_{\texttt{async}}(V)) \leq \mathcal{F}^n_{\texttt{async}}(V)$ and

$$V_{\lim} \leq \mathcal{F}^n_{\texttt{async}}(V) \leq \mathcal{F}^n(V).$$

Taking the limit as $n \to \infty$ gives us that $\lim_{n\to\infty} \mathcal{F}^n_{\texttt{async}}(V) = V_{\lim}$ if $\mathcal{F}(V) \leq V$. Using a symmetric argument, we get that $\lim_{n\to\infty} \mathcal{F}^n_{\texttt{async}}(V) = V_{\lim}$ if $\mathcal{F}(V) \geq V$.

Let $I \in \mathcal{V}$ be defined by $I(s,\sigma) = 1$ for all $(s,\sigma) \in S_1$. For a general $V \in \mathcal{V}$, we can find a $\delta > 0$ such that we have $V^- = V_{\lim} - \delta I \leq V \leq V_{\lim} + \delta I = V^+$ and $\mathcal{F}(V^-) \geq V^-$ and $\mathcal{F}(V^+) \leq V^+$. Therefore, using monotonicity of $\mathcal{F}_{\texttt{async}}$ we get

$$\mathcal{F}^n_{\texttt{async}}(V^-) \leq \mathcal{F}^n_{\texttt{async}}(V) \leq \mathcal{F}^n_{\texttt{async}}(V^+)$$

for all $n \geq 0$. Taking the limit as n tends to ∞ gives us the required result. Theorem 3.5 follows from a similar argument. □

A.7 Proof of Theorem 4.1

Given a function $Q : S_1 \times A \to \mathbb{R}$ we define a new function $\mathcal{H}(Q)$ using

$$\mathcal{H}(Q)(s,\sigma,a) = \bar{R}((s,\sigma),a) + \gamma \sum_{s' \in S} P(s' \mid s,a)[\![V_Q]\!](s',\sigma)$$

for all $(s,\sigma) \in S_1$ and $a \in A$. Then, Robust Option Q-learning is of the form

$$Q_{t+1}(s,\sigma,a) = (1-\alpha_t(s,\sigma,a))Q_t(s,\sigma,a) + \alpha_t(s,\sigma,a)\Big(H(Q_t)(s,\sigma,a) + w_t(s,\sigma,a)\Big)$$

where the noise factor is defined by

$$w_t(s,\sigma,a) = \gamma[\![V_{Q_t}]\!](\tilde{s},\sigma) - \gamma \sum_{s' \in S} P(s' \mid s,a)[\![V_Q]\!](s',\sigma)$$

with $\tilde{s} \sim P(\cdot \mid s,a)$ being the observed sample. Let $\mathcal{X}_t$ denote the measure space generated by the set of random vectors $\{Q_0, Q_1, \ldots, Q_t, w_0, \ldots, w_{t-1}, \alpha_0, \ldots, \alpha_t\}$. Then, for all $(s,\sigma) \in S_1$, $a \in A$ and $t \geq 0$, we have

$$\mathbb{E}[w_t(s,\sigma,a) \mid \mathcal{X}_t] = 0$$

and

$$\mathbb{E}[w_t^2(s,\sigma,a) \mid \mathcal{X}_t] \leq 4\gamma^2 \max_{s' \in S} \Big\{[\![V_{Q_t}]\!]^2(s',\sigma)\Big\} \leq 4\gamma^2 \max_{(s',\sigma') \in S_1, a' \in A} \Big\{Q_t^2(s',\sigma',a')\Big\}.$$

Furthermore, using Lemmas A.1 and A.2 and the definition of V_Q we can conclude that $\mathcal{H}$ is a contraction w.r.t the ℓ_∞-norm and Q^* is the unique fixed point of $\mathcal{H}$. Therefore, the random sequence of Q-functions $\{Q_t\}_{t\geq 0}$ satisfies all assumptions in Proposition 4.4 of [7] implying that $Q_t \to Q^*$ as $t \to \infty$ with probability 1. □

B Experimental Details

All experiments were run on a 48-core machine with 512GB of memory and 8 GPUs. In all approaches (ours and baselines) except for MADDPG, the policy consists of one fully-connected NN per subtask, each with two hidden layers. MADDPG consists of two policies, one for the agent and one for the adversary, each with two hidden layers. In the case of MADDPG, the subtask is encoded in the observation using a one-hot vector. All hyperparameters were computed by grid search over a small set of values.

Rooms environment. The hidden dimension used is 64 for all approaches except MADDPG for which we uses 128 dimensional hidden layers. For DAGGER, NAIVE and AROSAC we run SAC with Adam optimizer (learning rate of $\alpha = 0.02$), entropy weight $\beta = 0.05$, Polyac rate 0.005 and batch size of 100. In each iteration of AROSAC and DAGGER, SAC is run for $N = 10000$ steps. Similarly, ROSAC is run with Adam optimizer (learning rates $\alpha_\psi = \alpha_\theta = 0.02$), entropy weight $\beta = 0.05$, Polyac rate 0.005 and batch size of 300. The MADDPG baseline uses a learning rate of 0.0003 and batch size of 256.

F1/10th environment. The hidden dimension used is 64 for all approaches. For DAGGER, NAIVE and AROSAC we run SAC with Adam optimizer (learning rate of $\alpha = 0.0003$), entropy weight $\beta = 0.01$, Polyac rate 0.005 and batch size of 256. In each iteration of AROSAC and DAGGER, SAC is run for $N = 10000$ steps. Similarly, ROSAC is run with Adam optimizer (learning rates $\alpha_\psi = \alpha_\theta = 0.0003$), entropy weight $\beta = 0.01$, Polyac rate 0.005 and batch size of 5×256. The MADDPG baseline uses a learning rate of 0.0003 and batch size of 256.