Appendix A. Proof of Theorem 5

Proof First, we prove that $(TV)_A \overset{s}{=} V_A$ implies $V_A \overset{s}{=} \max_{f \in \mathcal{W}_A} V_A^{fR_B(f,V_B)}$ for a fixed V_B.

Suppose that $V_A \overset{s}{=} (TV)_A$. Given a policy $f \in \mathcal{W}_A$, let $f_s = f(s)$ for each $s \in \mathcal{S}$. Then, we obtain for all $s \in \mathcal{S}$

$$
\begin{aligned}
V_A(s) &= (TV)_A(s) && \text{(9a)}\\
&= \sup_{f_s \in \Delta(\mathcal{A})} r_A(s, f_s, R_B(s, f_s, V_B)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f_s, R_B(s,f_s,V_B))} \left[V_A(s')\right] && \text{(9b)}\\
&\geq r_A(s, f_s, R_B(s, f_s, V_B)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f_s, R_B(s,f_s,V_B))} \left[V_A(s')\right] && \text{(9c)}\\
&= r_A(s, f(s), R_B(s, f(s), V_B)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f(s), R_B(s,f(s),V_B))} \left[V_A(s')\right]. && \text{(9d)}
\end{aligned}
$$

By repeatedly applying this inequality, we obtain

$$
\begin{aligned}
V_A(s) &\geq r_A(s, f(s), R_B(s, f(s), V_B)) + \gamma_A \mathbb{E}_{s_1}\left[V_A(s_1)\right] && \text{(10a)}\\
&= r_A(s, f(s), R_B(s, f(s), V_B)) && \text{(10b)}\\
&\quad + \gamma_A \mathbb{E}_{s_1}\Big[r_A(s_1, f(s_1), R_B(s_1, f(s_1), V_B)) + \gamma_A \mathbb{E}_{s_2}\left[V_A(s_2)\right]\Big]\\
&\geq \cdots\\
&= \mathbb{E}^{fR_B(f,V_B)}\left[\sum_{t=0}^{n-1} \gamma_A^t \hat{r}_A(s_t, a_t, b_t) \middle| s_0 = s\right] + \gamma_A^n \mathbb{E}^{fR_B(f,V_B)}\Big[V_A(s_n)\Big| s_0 = s\Big]. && \text{(10c)}
\end{aligned}
$$

In the limit for $n \to \infty$, the first term of the right-hand side of Equation (10c) converges to $V_A^{fR_B(f,V_B)}(s)$, while the second term converges to 0[6]. As the choice of f is arbitrary, we obtain $V_A \succcurlyeq \sup_{f \in \mathcal{W}_A} V_A^{fR_B(f,V_B)}$.

We show $V_A \overset{s}{=} \max_{f \in \mathcal{W}_A} V_A^{fR_B(f,V_B)}$. It suffices to show that there exists a policy $f^* \in \mathcal{W}_A$ such that $V_A \overset{s}{=} V_A^{f^* R_B(f^*,V_B)}$. Because $R_A(s,V)$ is non-empty, let $f_s^* \in R_A(s,V)$ for each $s \in \mathcal{S}$ and f^* be a policy whose action distribution at s is $f^*(s) = f_s^*$ for each s, we have for all $s \in \mathcal{S}$

$$
\begin{aligned}
V_A(s) &= (TV)_A(s) && \text{(11a)}\\
&= \sup_{f_s \in \Delta(\mathcal{A})} r_A(s, f_s, R_B(s, f_s, V_B)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f_s, R_B(s,f_s,V_B))} \left[V_A(s')\right] && \text{(11b)}\\
&= r_A(s, f_s^*, R_B(s, f_s^*, V_B)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f_s^*, R_B(s,f_s^*,V_B))} \left[V_A(s')\right] && \text{(11c)}\\
&= r_A(s, f^*(s), R_B(s, f^*(s), V_B)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f^*(s), R_B(s,f^*(s),V_B))} \left[V_A(s')\right]. && \text{(11d)}
\end{aligned}
$$

Therefore, with the repeated application of this equality, analogously to the above argument, we obtain $V_A \overset{s}{=} V_A^{f^* R_B(f^*,V_B)}$.

Second, letting V_A be fixed, we prove that $(TV)_B \overset{s}{=} V_B$ implies $V_B \overset{s}{=} \max_{g \in \mathcal{W}_B} V_B^{fg}$ for $f = R_A(V)$.

6. Since $\hat{r}_A$ is bounded and $\gamma_A \in [0,1)$, the inside of the expectation of Equation (10c) is always bounded, which enables the exchange of the extreme and the expectation. Therefore, the same exchanges are performed in Equation (14c) and Equation (19c).

Suppose $V_B \overset{s}{=} (TV)_B$ holds. Then, by constructing R_B, we obtain

$$V_B(s) = (TV)_B(s) \geq r_B(s, R_A(s, V), b_s) + \gamma_B \mathbb{E}_{s' \sim p(s'|s, R_A(s,V), b_s)} \left[V_B(s') \right] \tag{12}$$

for all $s \in \mathcal{S}$ and for all $b_s \in \mathcal{B}$. It implies that

$$V_B(s) \geq r_B(s, R_A(s, V), g(s)) + \gamma_B \mathbb{E}_{s' \sim p(s'|s, R_A(s,V), g(s))} \left[V_B(s') \right] \tag{13}$$

for all $s \in \mathcal{S}$ and for all $g \in \mathcal{W}_B^d$. By recursively applying Equation (13) for V_B on the right-hand side, we have for all $s \in \mathcal{S}$

$$\begin{aligned}
V_B(s) &\geq r_A(s, R_A(s, V), g(s)) + \gamma_B \mathbb{E}_{s_1} \left[V_B(s_1) \right] && \text{(14a)} \\
&\geq r_A(s, R_A(s, V), g(s)) + \gamma_B \mathbb{E}_{s_1} \left[r_A(s_1, R_A(s_1, V), g(s_1)) + \gamma_B \mathbb{E}_{s_2} \left[V_B(s_2) \right] \right] && \text{(14b)} \\
&\geq \cdots \\
&\geq \mathbb{E}^{R_A(V)g} \left[\sum_{t=0}^{n-1} \gamma_B^t \hat{r}_B(s_t, a_t, b_t) \middle| s_0 = s \right] + \gamma_B^n \mathbb{E}^{R_A(V)g} \left[V_B(s_n) \middle| s_0 = s \right]. && \text{(14c)}
\end{aligned}$$

In the limit for $n \to \infty$, the first term of the right-hand side of Equation (14c) converges to $V_B^{R_A(V)g}(s)$, and the second term converges to 0. Therefore, $V_B(s) \geq V_B^{R_A(V)g}(s)$ holds for all $s \in \mathcal{S}$. Since $g \in \mathcal{W}_B^d$ is arbitrary, it holds that

$$V_B(s) \succcurlyeq \max_{g \in \mathcal{W}_B^d} V_B^{R_A(V)g}. \tag{15}$$

Owing to the existence of a deterministic optimal policy for any single-agent MDP, we obtain

$$V_B \succcurlyeq \max_{g \in \mathcal{W}_B^d} V_B^{R_A(V)g} \overset{s}{=} \max_{g \in \mathcal{W}_B} V_B^{R_A(V)g}. \tag{16}$$

We show that $V_B(s) \overset{s}{=} \max_{g \in \mathcal{W}_B} V_B^{R_A(V)g}$. It suffices to show that there exists g^* such that $V_B(s) \overset{s}{=} V_B^{R_A(V)g^*}$. Let g^* such that $g^*(s) = R_B(s, R_A(s, V), V_B)$ for each $s \in \mathcal{S}$. Then, by the definition of R_B, for any $s \in \mathcal{S}$, we have

$$V_B(s) = r_B(s, R_A(s, V), g^*(s)) + \gamma_B \mathbb{E}_{s' \sim p(s'|s, R_A(s,V), g^*(s))} \left[V_B(s') \right]. \tag{17}$$

With the repeated application of this equality, analogous to the above argument, we obtain $V_B \overset{s}{=} V_B^{R_A(V)g^*}$. ■

Appendix B. Proof of Theorem 7

Proof First, suppose that $V_A^* \in \mathcal{PV}$. Then, by the definition of the SE value function, we have $V_A^* \succcurlyeq v$ for all $v \in \mathrm{cl}\mathcal{V}$. Therefore, no other PO value function exists. Hence, $\mathcal{PV} = \{V_A^*\}$.

Subsequently, suppose that $\mathcal{PV} = \{v^*\}$ is a singleton. Then, $v^* \succ v$ for all $v \in \mathcal{V} \setminus \mathcal{PV}$ because, for all $v \in \mathcal{V} \setminus \mathcal{PV}$, there exists $\hat{v} \in \mathcal{PV}$ such that $\hat{v} \succ v$. Suppose that $v^* \overset{s}{\neq} V_A^*$.

Then, for some $s \in \mathcal{S}$, $v^*(s) < \sup_{f \in \mathcal{W}} V_A^{f\dagger}(s)$ holds. It implies that there exists $v \in \mathcal{V} \setminus \mathcal{PV}$ such that $v^*(s) < v(s)$ for some s, which contradicts $v^* \succ v$. Hence, $v^* \overset{s}{=} V_A^*$.

Altogether, we have $V_A^* \in \mathcal{PV}$ if and only if $\mathcal{PV}$ is a singleton. Because PO policies exists if and only if $\mathcal{PV} \cap \mathcal{V} \neq \emptyset$, the SE policy exists if and only if $\mathcal{PV} = \{v^*\}$ is a singleton and $v^* \in \mathcal{V}$. ■

Appendix C. Proof of Theorem 8

Proof We prove each statement one by one.

Proof of (a) The definition of $Q_A^{f'\dagger}(s, f)$ is equivalent to

$$Q_A^{f'\dagger}(s, f) = \mathbb{E}^{fR_B^*(f)}\left[\hat{r}_A(s, a, b)|s\right] + \gamma_A \mathbb{E}^{fR_B^*(f)}\left[V_A^{f'\dagger}(s')\middle|s\right]. \tag{18}$$

Then, if it holds that $V_A^{f'\dagger}(s) \leq Q_A^{f'\dagger}(s, f)$ for all $s \in \mathcal{S}$, we have, for all $s_0 \in \mathcal{S}$,

$$\begin{aligned}
V_A^{f'\dagger}(s_0) &\leq \mathbb{E}^{fR_B^*(f)}\left[\hat{r}_A(s_0, a_0, b_0)|s_0\right] + \gamma_A \mathbb{E}^{fR_B^*(f)}\left[V_A^{f'\dagger}(s_1)\middle|s_0\right] && \text{(19a)}\\
&\leq \mathbb{E}^{fR_B^*(f)}\left[\hat{r}_A(s_0, a_0, b_0)|s_0\right] + \gamma_A \mathbb{E}^{fR_B^*(f)}\Big[&& \text{(19b)}\\
&\qquad \mathbb{E}^{fR_B^*(f)}\left[\hat{r}_A(s_1, a_1, b_1)|s_1\right] + \gamma_A \mathbb{E}^{fR_B^*(f)}\left[V_A^{f'\dagger}(s_2)\middle|s_1\right]\Big|s_0\Big]\\
&\leq \dots\\
&\leq \mathbb{E}^{fR_B^*(f)}\left[\sum_{t=0}^{n-1} \gamma_A^t \hat{r}_A(s_t, a_t, b_t)\middle|s_0\right] + \gamma_A^n \mathbb{E}^{fR_B^*(f)}\left[V_A^{f'\dagger}(s_n)\middle|s_0\right]. && \text{(19c)}
\end{aligned}$$

In the limit for $n \to \infty$, the first term converges to $V_A^{fR_B^*(f)}(s_0)$ and the second term converges to 0. Therefore, it holds that $V_A^{f'\dagger}(s_0) \leq Q_A^{f'\dagger}(s, f) \leq V_A^{fR_B^*(f)}(s_0) = V_A^{f\dagger}(s_0)$ for all $s_0 \in \mathcal{S}$.

Proof of (b) If it holds that $Q_A^{f'\dagger}(s, f) = V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S}$, by the similar derivation of Equation (19), we obtain

$$V_A^{f'\dagger}(s_0) = \mathbb{E}^{fR_B^*(f)}\left[\sum_{t=0}^{n-1} \gamma_A^t \hat{r}_A(s_t, a_t, b_t)\middle|s_0\right] + \gamma_A^n \mathbb{E}^{fR_B^*(f)}\left[V_A^{f'\dagger}(s_n)\middle|s_0\right] \tag{20}$$

for all $s_0 \in \mathcal{S}$. Then, in the limit for $n \to \infty$, we have $V_A^{f'\dagger}(s_0) = V_A^{f\dagger}(s_0)$ for all $s_0 \in \mathcal{S}$.

Conversely, if it holds that $V_A^{f'\dagger}(s) = V_A^{f\dagger}(s)$ for all $s \in \mathcal{S}$, we have

$$\begin{aligned}
V_A^{f\dagger}(s) &= r_A(s, f(s), R_B^*(s, f)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f(s), R_B^*(s,f))}\left[V_A^{f\dagger}(s')\right] && \text{(21a)}\\
&= r_A(s, f(s), R_B^*(s, f)) + \gamma_A \mathbb{E}_{s' \sim p(s'|s, f(s), R_B^*(s,f))}\left[V_A^{f'\dagger}(s')\right] && \text{(21b)}\\
&= Q_A^{f'\dagger}(s, f) && \text{(21c)}
\end{aligned}$$

for all $s \in \mathcal{S}$, which follows that $Q_A^{f'\dagger}(s, f) = V_A^{f\dagger}(s) = V_A^{f'\dagger}(s)$ for all $s \in \mathcal{S}$.

Proof of (c) By the definition of $V_A^{f\dagger}(s)$, we can rewrite the reward value as

$$r_A(s, f(s), R_B^*(s, f)) = V_A^{f\dagger}(s) - \gamma_A \mathbb{E}_{s'}\left[V_A^{f\dagger}(s')\right]. \tag{22}$$

Plugging it into the definition of $Q_A^{f'\dagger}(s, f)$ and subtracting $V_A^{f'\dagger}(s)$, we have

$$\begin{aligned} Q_A^{f'\dagger}(s, f) - V_A^{f'\dagger}(s) &= r_A(s, f(s), R_B^*(s, f)) + \gamma_A \mathbb{E}_{s'}\left[V_A^{f'\dagger}(s')\right] - V_A^{f'}(s) && (23)\\ &= V_A^{f\dagger}(s) - V_A^{f'\dagger}(s) - \gamma_A \mathbb{E}_{s'}\left[V_A^{f\dagger}(s') - V_A^{f'\dagger}(s')\right]. && (24) \end{aligned}$$

The policy improvement is possible if and only if $Q_A^{f'\dagger}(\cdot, f) \succ V_A^{f'\dagger}$. Because of the above equality, equivalently, we can say that the policy improvement from f' is possible if and only if there exists a policy f whose value function satisfies

$$V_A^{f\dagger}(s) - V_A^{f'\dagger}(s) \geq \gamma_A \mathbb{E}_{s'}\left[V_A^{f\dagger}(s') - V_A^{f'\dagger}(s')\right] \tag{25}$$

for all $s \in \mathcal{S}$, and there exists a state s where the inequality strictly holds. ■

Appendix D. Proof of Theorem 9

Proof We prove the counterpart of Theorem 9. Taking the negation of Equation (5) for all $f \in \mathcal{W}_A$, we have

$$\begin{aligned} &\neg\forall f \in \mathcal{W}_A \left\{Q_A^{f'\dagger}(s, f) \geq V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S} \implies V_A^{f\dagger}(s) = V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S}\right\} \\ &\iff \neg\forall f \in \mathcal{W}_A \left\{Q_A^{f'\dagger}(s, f) < V_A^{f'\dagger}(s)\ \exists s \in \mathcal{S} \vee V_A^{f\dagger}(s) = V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S}\right\} && (26a)\\ &\iff \exists f \in \mathcal{W}_A \left\{Q_A^{f'\dagger}(s, f) \geq V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S} \wedge V_A^{f\dagger}(s) \neq V_A^{f'\dagger}(s)\ \exists s \in \mathcal{S}\right\}. && (26b) \end{aligned}$$

From Theorem 8 (a), since $V_A^{f\dagger}(s) \geq V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S}$ holds and $Q_A^{f'\dagger}(s, f) \geq V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S}$ holds, we have

$$\begin{aligned} &\exists f \in \mathcal{W}_A \left\{Q_A^{f'\dagger}(s, f) \geq V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S} \wedge V_A^{f\dagger}(s) \neq V_A^{f'\dagger}(s)\ \exists s \in \mathcal{S}\right\} \\ &\implies \exists f \in \mathcal{W}_A \left\{V_A^{f\dagger}(s) \geq V_A^{f'\dagger}(s)\ \forall s \in \mathcal{S} \wedge V_A^{f\dagger}(s) \neq V_A^{f'\dagger}(s)\ \exists s \in \mathcal{S}\right\} && (27)\\ &\iff \exists f \in \mathcal{W}_A \left\{V_A^{f\dagger} \succ V_A^{f'\dagger}\right\} && (28)\\ &\implies \exists v \in \mathcal{V} \subseteq \mathrm{cl}\mathcal{V} \left\{v \succ V_A^{f'\dagger}\right\} && (29)\\ &\implies V_A^{f'\dagger} \notin \mathcal{P}\mathcal{V}. && (30) \end{aligned}$$

Therefore, f' is not a PO policy. This completes the proof. ■

Appendix E. Proof of Theorem 10

Proof First, we prove that (i) $\implies V_A^{f'\dagger}(s) > V_A^{f\dagger}(s)\ \exists s \in \mathcal{S}$. Given $f \in \mathcal{W}_A, f' \in \mathcal{W}_A$, by the definition of $\delta(f, f')$, we have

$$\mathbb{E}_{s'\sim p(s'|s,f(s),R_B^*(s,f))}\left[Q_A^{f'\dagger}(s', f) - V_A^{f'\dagger}(s')\right] \leq \delta(f, f') \tag{31}$$

for all $s \in \mathcal{S}$. Thus, we have

$$\begin{aligned}
Q_A^{f'\dagger}(s, f) &= r_A(s, f(s), R_B^*(s, f)) + \gamma_A \mathbb{E}_{s'\sim p(s'|s,f(s),R_B^*(s,f))}\left[V_A^{f'\dagger}(s')\right] && \text{(32a)}\\
&\geq r_A(s, f(s), R_B^*(s, f)) && \text{(32b)}\\
&\quad + \gamma_A\Big(\mathbb{E}_{s'\sim p(s'|s,f(s),R_B^*(s,f))}\left[Q_A^{f'\dagger}(s', f)\right] - \delta(f, f')\Big)\\
&\geq \cdots\\
&\geq \mathbb{E}^{fR_B^*(f)}\left[\sum_{t=0}^{n-1}\gamma_A^t r_A(s_t, f(s_t), R_B^*(s_t, f)) + \gamma_A^n V_A^{f'\dagger}(s_n)\middle| s_0 = s\right] && \text{(32c)}\\
&\qquad - \left(\sum_{t=0}^{n-1}\gamma_A^t - \gamma_A^0\right)\delta(f, f').
\end{aligned}$$

for all $s \in \mathcal{S}$. In the limit for $n \to \infty$, the first term converges to $V_A^{f\dagger}(s)$, the second term converges to 0, and the third term converges to $-\frac{\gamma_A}{1-\gamma_A}\delta(f, f')$. Therefore, we have

$$Q_A^{f'\dagger}(s, f) \geq V_A^{f\dagger}(s) - \frac{\gamma_A}{1-\gamma_A}\delta(f, f') \tag{33}$$

for all $s \in \mathcal{S}$. Then, if (i) holds, because there exists $s \in \mathcal{S}$ such that $Q_A^{f'\dagger}(s, f) < V_A^{f'\dagger}(s) - \frac{\gamma_A}{1-\gamma_A}\delta(f, f')$, it holds that

$$V_A^{f'\dagger}(s) - \frac{\gamma_A}{1-\gamma_A}\delta(f,f') > Q_A^{f'\dagger}(s, f) \geq V_A^{f\dagger}(s) - \frac{\gamma_A}{1-\gamma_A}\delta(f, f') \quad \exists s \in \mathcal{S}. \tag{34}$$

$$\therefore V_A^{f'\dagger}(s) > V_A^{f\dagger}(s) \quad \exists s \in \mathcal{S}. \tag{35}$$

If (ii) holds, because of Theorem 8 (b), it holds that

$$V_A^{f'\dagger}(s) = V_A^{f\dagger}(s) \quad \forall s \in \mathcal{S}. \tag{36}$$

Therefore, it holds that

$$\begin{aligned}
&\forall f \in \mathcal{W}_A\{\text{(i) or (ii)}\}\\
&\quad\implies \forall f \in \mathcal{W}_A\left\{V_A^{f'\dagger}(s) > V_A^{f\dagger}(s)\ \ \exists s \in \mathcal{S}\ \text{ or }\ V_A^{f'\dagger}(s) = V_A^{f\dagger}(s)\ \ \forall s \in \mathcal{S}\right\},
\end{aligned} \tag{37}$$

where the right-hand side is equivalent to the definition of PO policies (Theorem 6). ■

Appendix F. Proof of Theorem 11

Proof We prove each statement one by one. The proof of (a) is based on the monotonicity of $\{v_t\}$ and the compactness of $\mathrm{cl}\mathcal{V}$. The proof of (b) is based on the statement (b) of Theorem 8. The statements (c) and (d) are proved by contradiction using the statement (c) of Theorem 8.

Proof of (a) As $f_{t+1} \in \mathcal{W}_{\succcurlyeq}(f_t)$, it holds that

$$v_{t+1} = V_A^{f_{f+1}\dagger} \succcurlyeq Q_A^{f_t\dagger}(\cdot, f_{t+1}) \succcurlyeq V_A^{f_t\dagger}(\cdot) = v_t. \tag{38}$$

Thus, $v_{t+1} \succcurlyeq v_t$ holds for all $t \in \mathbb{N}$.

Because r_A is bounded, so is $V_A^{f\dagger}$, implying $\mathrm{cl}\mathcal{V}$ is compact. Considering the sequence $\{v_t(s)\}_{t=0}^{\infty}$ for each $s \in \mathcal{S}$, it is a monotonically increasing sequence. $\{v_t(s)\}_{t=0}^{\infty}$ converges in $\mathrm{cl}\mathcal{V}$ because $\mathrm{cl}\mathcal{V}$ is compact. Let $v_\infty(s) = \lim_{t\to\infty} v_t(s)$ for each $s \in \mathcal{S}$. Then, it holds that $v_\infty = \lim_{t\to\infty} v_t$.

Proof of (b) First, $v_{t+1} \overset{s}{=} v_t \impliedby v_t \overset{s}{=} v_\infty$ is apparent.

Second, we prove $v_{t+1} \overset{s}{=} v_t \implies v_t \overset{s}{=} v_\infty$. As $f_{t+1} \in \mathcal{W}_{1-\epsilon}(f_t)$, it holds that

$$\mathcal{L}\left[V_A^{f_{t+1}\dagger}\right] - \mathcal{L}\left[V_A^{f_t\dagger}\right] \geq \mathcal{L}\left[Q_A^{f_t\dagger}(\cdot, f_{t+1})\right] - \mathcal{L}\left[V_A^{f_t\dagger}\right] \tag{39}$$

$$\geq (1-\epsilon) \sup_{f\in\mathcal{W}_{\succcurlyeq}(f_t)} \left(\mathcal{L}\left[Q_A^{f_t\dagger}(\cdot, f)\right] - \mathcal{L}\left[V_A^{f_t\dagger}\right]\right). \tag{40}$$

If $v_{t+1} \overset{s}{=} v_t$, we have $\mathcal{L}[v_{t+1}] - \mathcal{L}[v_t] = 0$; hence, the above inequality implies

$$\sup_{f\in\mathcal{W}_{\succcurlyeq}(f_t)} \mathcal{L}\left[Q_A^{f_t\dagger}(\cdot, f)\right] = \mathcal{L}\left[V_A^{f_t\dagger}\right]; \tag{41}$$

therefore, for all $f \in \mathcal{W}_{\succcurlyeq}(f_t)$, we have $\mathcal{L}\left[Q_A^{f_t\dagger}(\cdot, f)\right] = \mathcal{L}\left[V_A^{f_t\dagger}\right]$. Because $\mathcal{L}$ is Pareto-compliant, the above equality implies $Q_A^{f_t\dagger}(\cdot, f) \not\succ V_A^{f_t\dagger}$. However, because $f \in \mathcal{W}_{\succcurlyeq}(f_t)$, we have $Q_A^{f_t\dagger}(\cdot, f) \succcurlyeq V_A^{f_t\dagger}$, implying $Q_A^{f_t\dagger}(\cdot, f) \overset{s}{=} V_A^{f_t\dagger}$. From Statement (b) of Theorem 8, we have $V_A^{f\dagger} \overset{s}{=} V_A^{f_t\dagger}$ for all $f \in \mathcal{W}_{\succcurlyeq}(f_t)$. (Conversely, if $V_A^{f\dagger} \overset{s}{=} V_A^{f_t\dagger}$ for all $f \in \mathcal{W}_{\succcurlyeq}(f_t)$, we have $v_{t+1} \overset{s}{=} v_t$.) Thus, $\mathcal{W}_{\succcurlyeq}(f) = \mathcal{W}_{\succcurlyeq}(f_t)$ holds for all $f \in \mathcal{W}_{\succcurlyeq}(f_t)$. It implies that $\mathcal{W}_{\succcurlyeq}(f_{t+1}) = \mathcal{W}_{\succcurlyeq}(f_t)$ since $f_{t+1} \in \mathcal{W}_{\succcurlyeq}(f_t)$; thus, $f_{t+2} \in \mathcal{W}_{\succcurlyeq}(f_{t+1}) = \mathcal{W}_{\succcurlyeq}(f_t)$. Therefore,

$$V_A^{f_t\dagger} \overset{s}{=} V_A^{f_{t+1}\dagger} \overset{s}{=} V_A^{f_{t+2}\dagger} \overset{s}{=} \cdots \tag{42}$$

and hence $v_{t+k} \overset{s}{=} v_t$ for all $k \geq 0$, implying that $v_t = v_\infty$.

Proof of (c) First, we show that

$$\min_{s\in\mathcal{S}}(v(s) - v_\infty(s)) \leq \gamma_A(\max_{s\in\mathcal{S}}(v(s) - v_\infty(s)))$$

for all $v \in \mathcal{V}$.

Suppose that there exists $v^* \in \mathcal{V}$ such that

$$\min_{s\in\mathcal{S}}(v^*(s) - v_\infty(s)) > \gamma_A(\max_{s\in\mathcal{S}}(v^*(s) - v_\infty(s)))$$

holds. Then, $v^* \succ v_\infty$ and there exists $f^* \in \mathcal{W}_A$ whose value function is $V_A^{f^*\dagger} \overset{s}{=} v^*$. Let $\tau = \frac{\min_{s\in\mathcal{S}}(v^*(s)-v_\infty(s))}{\max_{s\in\mathcal{S}}(v^*(s)-v_\infty(s))}$. We have $1 \geq \tau > \gamma_A$.

For an arbitrarily small $\xi > 0$ there exists t such that $\|v_t - v_\infty\|_\infty \leq \xi$ because $v_\infty \overset{s}{=} \lim_{t\to\infty} v_t$. Because of the continuity of $\mathcal{L}$, it also implies that for an arbitrarily small $\xi' > 0$ there exists t' such that $\mathcal{L}[v_{t'}] > \mathcal{L}[v_\infty] - \xi'$. Therefore, we can find t satisfying

$$\|v_\infty - v_t\|_\infty \leq \delta\frac{\tau-\gamma_A}{\gamma_A}\|v^* - v_\infty\|_\infty \quad \text{for some } \delta \in (0,1), \quad \text{and} \tag{43}$$

$$\mathcal{L}[v_\infty] - \mathcal{L}[v_t] < (1-\epsilon)(\mathcal{L}[\bar{v}] - \mathcal{L}[v_\infty]), \tag{44}$$

where $\bar{v}(s) := v_\infty(s) + (1-\delta)(\tau-\gamma_A)\|v^* - v_\infty\|_\infty$. Let f_t be the corresponding policy whose value function is $V_A^{f_t\dagger} \overset{s}{=} v_t$. Here, we assume that $\gamma_A > 0$. For the case of $\gamma_A = 0$, Statement (c) is an immediate consequence of Statement (d) because $v_\infty \in \mathcal{PV}$ and $\mathcal{PV} \subseteq \partial\mathcal{V}$.

Subsequently, we show $V_A^{f^*\dagger} \succcurlyeq Q_A^{f_t\dagger}(\cdot, f^*) \succcurlyeq \bar{v}_\infty$ using Equation (43). In the proof of Statement (c) of Theorem 8, for all $s \in \mathcal{S}$,

$$Q_A^{f'\dagger}(s,f) - V_A^{f'\dagger}(s) = V_A^{f\dagger}(s) - V_A^{f'\dagger}(s) - \gamma_A\mathbb{E}_{s'}\left[V_A^{f\dagger}(s') - V_A^{f'\dagger}(s')\right]. \tag{45}$$

Notably, $\mathbb{E}_{s'}\left[V_A^{f\dagger}(s') - V_A^{f'\dagger}(s')\right] \leq \|V_A^{f\dagger} - V_A^{f'\dagger}\|_\infty$. Letting $f = f^*$ and $f' = f_t$, we have, for all $s \in \mathcal{S}$,

$$Q_A^{f_t\dagger}(s,f^*) - V_A^{f_t\dagger}(s) \geq V_A^{f^*\dagger}(s) - V_A^{f_t\dagger}(s) - \gamma_A\|V_A^{f^*\dagger} - V_A^{f_t\dagger}\|_\infty. \tag{46}$$

By adding $V_A^{f_t\dagger}(s) - v_\infty(s)$ to both sides of the above inequality, we obtain, for all $s \in \mathcal{S}$,

$$\begin{aligned}
Q_A^{f_t\dagger}(s,f^*) - v_\infty(s) &\geq v^*(s) - v_\infty(s) - \gamma_A\|v^* - v_t\|_\infty && \text{(47a)}\\
&\geq \min_{s\in\mathcal{S}}(v^*(s) - v_\infty(s)) - \gamma_A\|v^* - v_t\|_\infty && \text{(47b)}\\
&= \min_{s\in\mathcal{S}}(v^*(s) - v_\infty(s)) - \gamma_A\|v^* - v_\infty + v_\infty - v_t\|_\infty && \text{(47c)}\\
&\geq \min_{s\in\mathcal{S}}(v^*(s) - v_\infty(s)) - \gamma_A\|v^* - v_\infty\|_\infty - \gamma_A\|v_\infty - v_t\|_\infty && \text{(47d)}\\
&= (\tau-\gamma_A)\|v^* - v_\infty\|_\infty - \gamma_A\|v_\infty - v_t\|_\infty && \text{(47e)}\\
&\geq (1-\delta)(\tau-\gamma_A)\|v^* - v_\infty\|_\infty. && \text{(47f)}
\end{aligned}$$

Therefore, we have $Q_A^{f_t\dagger}(\cdot, f^*) \succcurlyeq \bar{v}$. Because $\bar{v} \succ v_\infty \succcurlyeq V_A^{f_t\dagger}$, it implies that $Q_A^{f_t\dagger}(\cdot, f^*) \succcurlyeq V_A^{f_t\dagger}$. From Statement (a) of Theorem 8, we have $V_A^{f^*\dagger} \succcurlyeq Q_A^{f_t\dagger}(\cdot, f^*)$.

Finally, we derive a contradiction. Let $\ell_t^{\sup} = \sup_{f\in\mathcal{W}_\succcurlyeq(f_t)}\mathcal{L}[Q_A^{f_t\dagger}(\cdot,f)]$. Because $f^* \in \mathcal{W}_\succcurlyeq(f_t)$, we have $\ell_t^{\sup} \geq \mathcal{L}[Q_A^{f_t\dagger}(\cdot,f^*)]$. We also know that $\mathcal{L}[Q_A^{f_t\dagger}(\cdot,f^*)] \geq \mathcal{L}[\bar{v}]$. Because $f_{t+1} \in \mathcal{W}_{1-\epsilon}(f_t)$, we obtain

$$\begin{aligned}
\mathcal{L}[v_{t+1}] - \mathcal{L}[v_t] &\geq \mathcal{L}[Q_A^{f_t\dagger}(\cdot,f_{t+1})] - \mathcal{L}[v_t] && \text{(48a)}\\
&\geq (1-\epsilon)\left(\sup_{f\in\mathcal{W}_\succcurlyeq(f_t)}\mathcal{L}[Q_A^{f_t\dagger}(\cdot,f)] - \mathcal{L}[v_t]\right) && \text{(48b)}\\
&\geq (1-\epsilon)(\mathcal{L}[Q_A^{f_t\dagger}(\cdot,f^*)] - \mathcal{L}[v_\infty]) && \text{(48c)}\\
&\geq (1-\epsilon)(\mathcal{L}[\bar{v}] - \mathcal{L}[v_\infty]) && \text{(48d)}\\
&> \mathcal{L}[v_\infty] - \mathcal{L}[v_t], && \text{(48e)}
\end{aligned}$$

where we used Equation (44) for the last inequality. It implies $\mathcal{L}[v_{t+1}] > \mathcal{L}[v_\infty]$, which contradicts to the fact that $v_\infty \succcurlyeq v_{t+1}$. Therefore, we have $\min_{s\in\mathcal{S}}(v(s) - v_\infty(s)) \leq \gamma_A \left(\max_{s\in\mathcal{S}}(v(s) - v_\infty(s))\right)$ for all $v \in \mathcal{V}$.

Subsequently, we show that $v_\infty \in \partial\mathcal{V}$.

Suppose that $v_\infty \in \mathcal{V} \setminus \partial\mathcal{V}$ (i.e., v_∞ is an interior of $\mathcal{V}$). Then, there exists an $r > 0$ such that $\mathcal{E} = \{v \in \mathcal{F}_\mathcal{S} \mid \|v - v_\infty\|_\infty \leq r\} \subseteq \mathcal{V}$. Let v^r_∞ be such that $v^r_\infty(s) = v_\infty(s) + r$ for all $s \in \mathcal{S}$. Then, $v^r_\infty \in \mathcal{E}$. Moreover, because $v^r_\infty \in \mathcal{V}$, there exists a policy f^r_∞ such that $V_A^{f^r_\infty\dagger} = v^r_\infty$.

We lead to the contradiction to $\min_{s\in\mathcal{S}} v(s) - v_\infty(s) \leq \gamma_A \left(\max_{s\in\mathcal{S}} v(s) - v_\infty(s)\right)$. Using Equation (43), we obtain

$$\frac{\min_{s\in\mathcal{S}} V_A^{f^r_\infty\dagger}(s) - v_\infty(s)}{\max_{s\in\mathcal{S}} V_A^{f^r_\infty\dagger}(s) - v_\infty(s)} = \frac{\min_{s\in\mathcal{S}}(v^r_\infty(s) - v_\infty(s))}{\max_{s\in\mathcal{S}}(v^r_\infty(s) - v_\infty(s))} = 1, \tag{49}$$

contradicting $\min_{s\in\mathcal{S}} v(s) - v_\infty(s) \leq \gamma_A \left(\max_{s\in\mathcal{S}} v(s) - v_\infty(s)\right)$ for all $v \in \mathcal{V}$.

Proof of (d) In case of $\gamma_A = 0$, we have $Q_A^{f_t\dagger}(\cdot, f) \overset{s}{=} V_A^{f\dagger}$ for all $f_t, f \in \mathcal{W}_A$.

Suppose that $v_\infty \notin \mathcal{PV}$. Then, there exists a value function $v^* \in \mathrm{cl}\mathcal{V}$ such that $v^* \succ v_\infty$. Similar to the proof of Statement (c), we choose t such that

$$\mathcal{L}[v_\infty] - \mathcal{L}[v_t] < (1 - \epsilon)(\mathcal{L}[v^*] - \mathcal{L}[v_\infty]). \tag{50}$$

Let $\mathcal{V}_{\succcurlyeq}(v_t) := \{v \in \mathcal{V} \mid v \succcurlyeq v_t\}$ and $\mathrm{cl}\mathcal{V}_{\succcurlyeq}(v_t) := \{v \in \mathrm{cl}\mathcal{V} \mid v \succcurlyeq v_t\}$. Then, we have $v^* \in \mathrm{cl}\mathcal{V}_{\succcurlyeq}(v_t)$ because $v^* \succ v_\infty \succcurlyeq v_t$. Moreover, because $\mathcal{V}_{\succcurlyeq}(v_t) = \{V_A^{f\dagger} : f \in \mathcal{W}_{\succcurlyeq}(f_t)\}$, we obtain

$$\begin{aligned}
\ell_t^{\sup} &:= \sup_{f\in\mathcal{W}_{\succcurlyeq}(f_t)} \mathcal{L}[V_A^{f\dagger}] && \text{(51a)}\\
&= \sup_{v\in\mathcal{V}_{\succcurlyeq}(v_t)} \mathcal{L}[v] && \text{(51b)}\\
&= \max_{v\in\mathrm{cl}\mathcal{V}_{\succcurlyeq}(v_t)} \mathcal{L}[v] && \text{(51c)}\\
&\geq \mathcal{L}[v^*]. && \text{(51d)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{L}[v_{t+1}] - \mathcal{L}[v_t] &\geq (1 - \epsilon)(\ell_t^{\sup} - \mathcal{L}[v_t]) && \text{(52a)}\\
&\geq (1 - \epsilon)(\ell_t^{\sup} - \mathcal{L}[v_\infty]) && \text{(52b)}\\
&\geq (1 - \epsilon)(\mathcal{L}[v^*] - \mathcal{L}[v_\infty]) && \text{(52c)}\\
&> \mathcal{L}[v_\infty] - \mathcal{L}[v_t], && \text{(52d)}
\end{aligned}$$

where the derivation of Equation (52a) is due to the definition of $\mathcal{W}_{1-\epsilon}(f_t)$ in Equation (7) and $Q_A^{f_t\dagger}(\cdot, f) \overset{s}{=} V_A^{f\dagger}$. This implies $\mathcal{L}[v_{t+1}] > \mathcal{L}[v_\infty]$, which contradicts to $v_\infty \succcurlyeq v_{t+1}$. Therefore, $v_\infty \in \mathcal{PV}$. ■

Appendix G. Splitting the Search Space

The proposed algorithm in Algorithm 1 can be impractical when the leader's policy space $\mathcal{W}_A$ is prohibitively large to search for the next policy in Equation (6) and Equation (8). To avoid the exhaustive search, we propose a splitting strategy for the search space to obtain improved policies from each subspace efficiently.

We first split $\mathcal{W}_A$ into disjoint sets $\mathcal{W}_1 \cup \cdots \cup \mathcal{W}_K$, where the best response is $R_B^*(f) = g_i^* \in \mathcal{W}_B^d$ for $f \in \mathcal{W}_i$. Because the set of deterministic policies, $\mathcal{W}_B^d$, is finite, we can always find such a separation. Additionally, we argue that the number of deterministic policies that form the best response is rather limited, enabling us to enumerate them easily.

We consider replacing (8) with locating a Pareto optimal f with respect to $Q_A^{f_t\dagger}(\cdot, f)$. Because (8) also locates Pareto optimal f, this replacement corresponds to selecting a different $\mathcal{L}$ at each t.

This is realized in two steps. First, we locate a policy satisfying $Q_A^{f_t\dagger}(\cdot, f) \succcurlyeq V_A^{f_t\dagger}$. It is nontrivial as the best response R_B^* changes as f changes. However, if we limit our attention to a subset $\mathcal{W}_k$ for some k, the best response does not change with $f \in \mathcal{W}_k$. Then, locating $f \in \mathcal{W}_k$ satisfying $Q_A^{f_t\dagger}(\cdot, f) \succcurlyeq V_A^{f_t\dagger}$ is casted as the following constrained optimization problem:

$$\min_{z, f \in \mathcal{W}_k} z \quad \text{s.t.} \quad V_A^{f_t\dagger}(s) - Q_A^{f_t\dagger}(s, f) \leqslant z \quad \text{for all } s \in \mathcal{S}, \tag{53}$$

where $Q_A^{f_t\dagger}(s, f)$ is differentiable with respect to f and is easily computable as the best response is constant. Therefore, it will be solved by, e.g., a gradient-based solver with projection onto $\mathcal{W}_k$. A solution with $z \leq 0$ satisfies $Q_A^{f_t\dagger}(\cdot, f) \succcurlyeq V_A^{f_t\dagger}$. Such a solution must exist in $\mathcal{W}_k$ for some k because f_t satisfies this condition.

Once we find such a solution, denoted as (z_t, f_t'), we find a Pareto optimal policy for $Q_A^{f_t\dagger}(\cdot, f)$ in $\mathcal{W}_k \cap \mathcal{W}_{\succcurlyeq}(f_t)$ by improving f_t'. It is done by performing a Pareto ascent method with projection,

$$f \leftarrow \Pi_{\mathcal{W}_k}(f + \eta\Delta), \quad \text{where} \quad \Delta^\top \nabla Q_A^{f_t\dagger}(s, f) \geqslant 0 \quad \text{for all } s \in \mathcal{S}, \tag{54}$$

or its variants, e.g., Harada et al. (2006), with f_t as a feasible initial solution of this step. If we find f_t^* such that $Q_A^{f_t\dagger}(\cdot, f) \succ V_A^{f_t\dagger}$, we let $f_{t+1} = f_t^*$. These steps will be continued until a predefined termination criterion is satisfied, or no policy strictly dominating the current policy is found. The whole algorithm is shown in Algorithm 2.

Algorithm 2 Practical Pareto-Optimal Policy Iteration

Input: Maximum number of iterations M.

1: Split the leader's policy space $\mathcal{W}_A$ into disjoint sets $\mathcal{W}_1 \cup \cdots \cup \mathcal{W}_K$, where the best response is $R_B^*(f) = g_i^* \in \mathcal{W}_B^d$ for $f \in \mathcal{W}_i$
2: Randomly initialize $f_0 \in \mathcal{W}_A$ and compute $g_0 = R_B^*(f_0)$
3: **for** $t = 0$ to $M - 1$ **do**
4: Compute $V_A^{f_t\dagger}$ by repeatedly applying $T_A^{f_t R_B^*(f_t)}$
5: Let $\mathcal{I}_t = \{1, \ldots, K\}$
6: **while** $|\mathcal{I}_t| > 0$ **do**
7: Randomly select $k \in \mathcal{I}_t$
8: Solve (53) to obtain z_t, f_t'
9: **if** $z_t \leq 0$ **then**
10: Perform the Pareto ascent (54) to obtain f_t^*
11: Let $f_{t+1} = f_t^*$ and **break if** $Q_A^{f_t\dagger}(\cdot, f_t^*) \succ V_A^{f_t\dagger}$
12: **end if**
13: Update $\mathcal{I}_t \leftarrow \mathcal{I}_t \setminus \{k\}$
14: **end while**
15: Let $f_M = \cdots = f_{t+1} = f_t$ and **break if** $\mathcal{I}_t = \emptyset$
16: **end for**

Output: A stationary policy f_M

Appendix H. Backtracking

Algorithm 1 may not output a PO policy because the terminate condition in line 5 of Algorithm 1 is the necessary condition for PO policies. In this case, Theorem 10 can be used to determine whether the output policy is a PO policy: provided (i) or (ii) in Theorem 10 holds for all $f \in \mathcal{W}_A$ with the output policy f_t and its state value function $V_A^{f_t\dagger}$, f_t is a PO policy. Notably, this judgment is not perfect because the policy may be determined not to be a PO policy when it really is a PO policy.

We proposed selecting another policy $f' \in \mathcal{W}(f_{t-1})$ that has not been selected thus far and restarting the for loop with $f_t \leftarrow f'$, which we call *"Backtracking"*, to deal with the case where the output policy $f_t (t < M)$ is determined not to be a PO policy. If there is no policy left in $\mathcal{W}(f_{t-1})$, then it backtracks one by one, as $\mathcal{W}(f_{t-2}), \mathcal{W}(f_{t-3}), \ldots$, until it reaches the set with policies that have never been selected. This method requires that $\mathcal{W}(f_t)$ is saved at each iteration.

The entire proposed algorithm with the PO-policy judgement and the backtracking is shown in Algorithm 3. W is a list to save each $\mathcal{W}(f_t)$ and M is the maximum number of iterations. This algorithm is designed such that the number of times to compute $V_A^{f_t\dagger}$ in line 5 and $\mathcal{W}_{\succcurlyeq}(f_t)$ in line 7 is at most M. If the PO-policy judgment in line 9 does not return `True` during M iterations, the algorithm outputs the policy that maximizes the objective $\sum_{s\in\mathcal{S}} \alpha_s V_A^{f_t\dagger}(s)$ among the ones that satisfy the necessary condition for PO policies thus far and f_M (see lines 12, 25, and 26).

Algorithm 3 Pareto-Optimal Policy Iteration with Backtracking

Input: Pareto-compliant scalarization $\mathcal{L}$, maximum number of iterations M, a M-length list of empty sets W.

$W(0) \leftarrow \mathcal{W}_A$
Randomly sample f_0 from $W(0)$ without replacement.
$\mathcal{W}_{output} \leftarrow \{\}$
for $t = 0$ to $M-1$ **do**
  Compute $V_A^{f_t\dagger} \in \mathcal{F}(\mathcal{S})$ such that $(T_A^{f_t R_B^*(f_t)} V_A^{f_t\dagger})(s) = V_A^{f_t\dagger}(s)$ for all $s \in \mathcal{S}$.
  $Q_A^{f_t\dagger}(s,a,b) \leftarrow r_A(s,a,b) + \gamma_A \mathbb{E}_{s'\sim p(s'|s,a,b)}\left[V_A^{f_t\dagger}(s')\right]$ for all $s \in \mathcal{S}, a \in \mathcal{A}, b \in \mathcal{B}$.
  $\mathcal{W}_{\succcurlyeq}(f_t) \leftarrow \{f \in \mathcal{W}_A \mid \mathbb{E}_{a\sim f(s)}\left[Q_A^{f_t\dagger}(s,a,R_B^*(s,f))\right] \geq V_A^{f_t\dagger}(s)\ \forall s \in \mathcal{S}\}$.
  if $\mathbb{E}_{a\sim f(s)}\left[Q_A^{f_t\dagger}(s,a,R_B^*(s,f))\right] = V_A^{f_t\dagger}(s)\ \forall s \in \mathcal{S}, \forall f \in \mathcal{W}_{\succcurlyeq}(f_t)$ **then**
    if f_t satisfies (i) or (ii) of Theorem 10 for all $f \in \mathcal{W}_A$ **then**
      return $f^* \leftarrow f_t$.
    end if
    Put f_t into $\mathcal{W}_{output}$.
    # Backtracking
    $k \leftarrow t$
    while $|W(k)| = 0$ **do**
      $k \leftarrow k-1$
    end while
    Randomly sample f_{t+1} from $W(k)$ without replacement.
  else
    $\mathcal{W}(f_t) \leftarrow \operatorname{argmax}_{f\in\mathcal{W}_{\succcurlyeq}(f_t)} \mathcal{L}\left[\mathbb{E}_{a\sim f(\cdot)}\left[Q_A^{f_t\dagger}(\cdot,a,R_B^*(\cdot,f))\right]\right]$.
    $W(t+1) \leftarrow \mathcal{W}(f_t)$.
    Randomly sample f_{t+1} from $W(t+1)$ without replacement.
  end if
end for
Put f_M into $\mathcal{W}_{output}$.
return $f^* \in \operatorname{argmax}_{f_t\in\mathcal{W}_{output}} \mathcal{L}\left[V_A^{f_t\dagger}(\cdot)\right]$

Output: A stationary policy f^*

Appendix I. Relation between the methods of Zhang et al. (2020) and Bucarey et al. (2022)

Zhang et al. (2020a) defined the optimal value functions as $V_i^*(i \in \{A,B\})$, which are the solution of the following equations (Zhang et al., 2020a): let $a \in \mathcal{A}, b \in \mathcal{B}, \gamma \in [0,1)$, and for all $s \in \mathcal{S}$,

$$Q_i^*(s,a,b) = r_i(s,a,b) + \gamma \mathbb{E}_{s'\sim p(s'|s,a,b)}\left[V_i^*(s')\right], \tag{55}$$

$$V_i^*(s) = \textsc{Stackelberg}_i(Q_A^*(s), Q_B^*(s)), \tag{56}$$

where $Q_i^*(s)$ is the Q-value $Q_i(a,b,s)$ of an action pair $(a,b) \in \mathcal{A}\times\mathcal{B}$ in state s. Given the fixed s, $Q_i^*(b,a,s)$ can be viewed as the payoff of (a,b) for the agent i; thus, $Q_i^*(s)$ can be

viewed as the payoff table of agent i in the state s. $\textsc{Stackelberg}_i(Q_A^*(s), Q_B^*(s))$ is the payoff for agent i in the Stackelberg equilibrium of the normal-form game represented by the payoff tables $(Q_A^*(s), Q_B^*(s))$. Therefore, Equation (56) can be written as

$$V_i^*(s) = Q_i^*(s, R'_A(s), R'_B(s, R'_A(s))), \tag{57a}$$
$$\text{where } R'_A(s) := \operatorname*{argmax}_{a\in\mathcal{A}} Q_A^*(s, a, R_B(s, a)), \tag{57b}$$
$$R'_B(s, a) := \operatorname*{argmax}_{b\in\mathcal{B}} Q_B^*(s, a, b). \tag{57c}$$

By substituting Equation (55) to Equation (57), we obtain

$$V_i^*(s) = r_i(s, R'_A(s), R'_B(s, R'_A(s))) + \gamma\mathbb{E}_{s'\sim p(s'|s,R'_A(s),R_B(s,R'_A(s)))}\left[V_i^*(s')\right], \tag{58a}$$
$$\text{where } R'_A(s) := \operatorname*{argmax}_{a\in\mathcal{A}} r_A(s, a, R'_B(s, a)) + \gamma\mathbb{E}_{s'\sim p(s'|s,a,R_B(s,a))}\left[V_A^*(s')\right], \tag{58b}$$
$$R'_B(s, a) := \operatorname*{argmax}_{b\in\mathcal{B}} r_B(s, a, b) + \gamma\mathbb{E}_{s'\sim p(s'|s,a,b))}\left[V_B^*(s')\right]. \tag{58c}$$

Meanwhile, the operator used in Bucarey et al. (2022) is given by

$$(Tv)_i(s) = r_i(s, R_A(s, v), R_B(s, R_A(s, v), v_B)) + \gamma\mathbb{E}_{s'\sim p(s'|s,R_A(s,v),R_B(s,R_A(s,v),v_b))}\left[v_i(s')\right], \tag{59a}$$
$$\text{where } R_A(s, v) := \operatorname*{argmax}_{f_s\in\Delta(\mathcal{A})} r_A(s, f_s, R_B(s, f_s, v_B)) + \gamma\mathbb{E}_{s'\sim p(s'|s,f_s,R_B(s,f_s,v_B))}\left[v_A(s')\right], \tag{59b}$$
$$R_B(s, f_s, v_B) := \operatorname*{argmax}_{b\in\mathcal{B}} r_B(s, f_s, b) + \gamma\mathbb{E}_{s'\sim p(s'|s,f_s,b)}\left[v_B(s')\right]. \tag{59c}$$

Notably, $\gamma = \gamma_A = \gamma_B$. Let us consider only the deterministic leader policies in the fixed point of Equation (59) and replace $\operatorname{argmax}_{f_s\in\Delta(\mathcal{A})}$ with $\operatorname{argmax}_{a\in\mathcal{A}}$ in the definition of R_A. Then, letting (V_A, V_B) be the fixed point of T, it holds from Equation (59), for all $s \in \mathcal{S}$, which is expressed as

$$V_i(s) = r_i(s, R_A(s, V), R_B(s, R_A(s, V), V_B)) + \gamma\mathbb{E}_{s'\sim p(s'|s,R_A(s,V),R_B(s,R_A(s,V),V_B))}\left[V_i(s')\right], \tag{60a}$$
$$\text{where } R_A(s, V) := \operatorname*{argmax}_{a\in\mathcal{A}} r_A(s, a, R_B(s, a, V_B)) + \gamma\mathbb{E}_{s'\sim p(s'|s,a,R_B(s,a,V_B))}\left[V_A(s')\right], \tag{60b}$$
$$R_B(s, a, V_B) := \operatorname*{argmax}_{b\in\mathcal{B}} r_B(s, a, b) + \gamma\mathbb{E}_{s'\sim p(s'|s,a,b)}\left[V_B(s')\right]. \tag{60c}$$

Comparing these with Equation (58), the definition of (V_A^*, V_B^*) by Equation (58) is equivalent to the definition of (V_A, V_B) by Equation (60). It follows that the fixed point in Zhang et al. (2020) is the same in Bucarey et al. (2022) when the leader policies in the fixed point are restricted to deterministic stationary policies.

By applying the discussion in the proof of Theorem 5 to (V_A, V_B) defined by Equation (60), we obtain

$$V_A(s) = \max_{f\in\mathcal{W}_A^d} V_A^{fR_B(f,V_B)}(s), \quad V_B(s) = \max_{g\in\mathcal{W}_B} V_B^{R_A(V)g}(s), \tag{61}$$

for all $s \in \mathcal{S}$. Then, the problem of the method of Bucarey et al. discussed in Section 5 also occurs to (V_A, V_B): V_A does not coincide with the state value function of the deterministic SSE policies since $R_B(f, V_B)$ is not guaranteed to be the best response against any $f \in \mathcal{W}_A^d$.

Appendix J. Application: Policy Teaching by Intervention to the Transitions

In single-agent MDPs, the policy aimed by the agent can be changed by making changes to the agent's observations, reward signals, and transition probabilities, among others. For example, in an MDP $(\mathcal{S}, \mathcal{A}, p, \rho, r, \gamma)$, the changes in the observations and the transition probabilities are represented by replacement to another transition function p', which changes the agent's aim from the optimal policies of $(\mathcal{S}, \mathcal{A}, p, \rho, r, \gamma)$ to those of $(\mathcal{S}, \mathcal{A}, p', \rho, r, \gamma)$. This is the theoretical foundation of poisoning attacks against RL (Behzadan and Munir, 2017; Huang and Zhu, 2019; Zhang et al., 2020b; Rakhsha et al., 2020; Sun et al., 2020) and policy teaching (Zhang and Parkes, 2008; Zhang et al., 2009), which aims to guide the agent's policy learning.

SSGs can be viewed as a model of intervention in transition probabilities by the leader, where the follower is the victim of the attack or the teaching. Even if the follower takes the same action at each state, the transition can be shifted by the leader changing its actions because the transition probability depends on the actions of all the agents. Therefore, suppose the follower cannot observe both the attendance and the actions of the leader, then the follower's learning can be represented as learning the optimal policies of the single-agent MDP $(\mathcal{S}, \mathcal{A}_B, p^f, \rho, \hat{r}_B, \gamma_B)$, where $f \in \mathcal{W}_A$ is the leader's policy, $p^f(s'|s, b) := \mathbb{E}_{a \sim f(s)}[\hat{p}(s'|s, a, b)]$, and $\hat{r}_B : \mathcal{S} \times \mathcal{B} \to \mathbb{R}$. Notably, $R_B^*(f)$ is its optimal policy. Then, the problem of guiding the follower's aim to the leader's objective is expressed as

$$\max_{f \in \mathcal{W}_A} \mathbb{E}^{f R_B^*(f)} \left[\sum_{t=0}^{\infty} \gamma_A^t \hat{r}_A(s_t, a_t, b_t) \right], \tag{62}$$

where $\hat{r}_A$ is the leader's reward function that represents the desired follower's behavior intended by the leader. f^* is the optimal solution of Equation (62), and the state-action sequences generated by the corresponding follower's optimal policy $R_B^*(f)$ maximize not only the follower's expected cumulative discounted reward but also that of the leader's. This can be viewed as the follower attaining the leader's desired behavior when the follower achieves its own learning goal.

The optimal solution of Equation (62) is given by the PO policy f that maximizes $\mathcal{L}[V_A^{f\dagger}] = \sum_{s \in \mathcal{S}} \rho(s) V_A^{f\dagger}(s)$. Therefore, our proposed approach can be applied to this problem when the transition function is known and the follower's best response can be computed. Although the optimality of the obtained solution is not guaranteed unless $\gamma_A = 0$, our analysis guarantees the monotone improvement of the leader's policy, which is of great importance in practice. In contrast, existing methods trying to obtain the SSE policy may be inadequate because it is usual that the follower is not myopic, and hence, the SSE policy does not necessarily exist.