Supplementary Material

Contents

Achievement Speed Results from DO-HJ-PPO Experiments

Here we present additional results for RAA and RR problems solved with **DO-HJ-PPO**. In both settings, DO-HJ-PPO out-performs or matches the best of baselines with less tuning and faster arrival. Notably as the difficulty of the problem increases the gap increases significantly with DO-HJ-PPO remaining the sole algorithm that can achieve the task in reasonable time and in both RAA and RR categories.

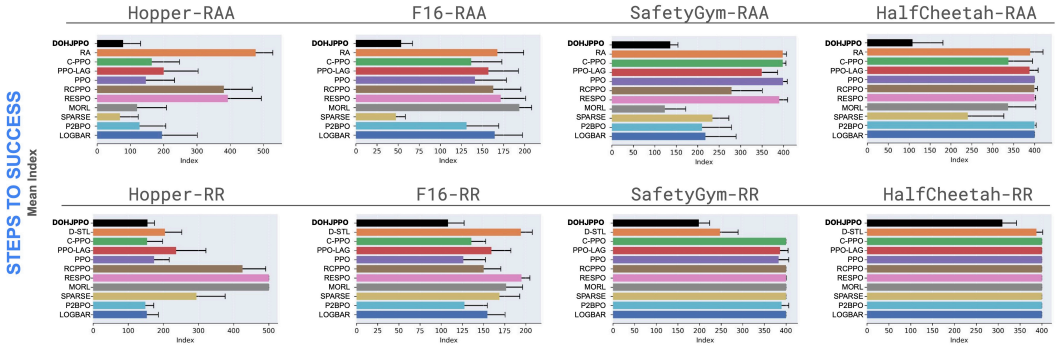

Figure 5: **Steps to Success (←) in RAA and RR Tasks for DO-HJ-PPO and Baselines** For the same 1000 trajectories in Figure 4, we quantify here the number of steps until achievement of both tasks: reaching without crash afterward in the RAA, reaching both goal in the RR. DO-HJ-PPO is not only competitive but consistently achieves the dual-objective problems in the fewest number of steps.

Proof Notation

Throughout the theoretical sections of this supplement, we use the following notation.

We let $\mathbb{N} = \{0, 1, \dots\}$ be the set of whole numbers.

We let $\mathbb{A}$ be the set of maps from $\mathbb{N}$ to $\mathcal{A}$. In other words, $\mathbb{A}$ is the set of sequences of actions the agent can choose. Given $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{A}$, and $\tau \in \mathbb{N}$, we let $[\mathbf{a}_1, \mathbf{a}_2]_\tau$ be the element of $\mathbb{A}$ for which

$$[\mathbf{a}_1, \mathbf{a}_2]_\tau(t) = \begin{cases} \mathbf{a}_1(t) & t < \tau, \\ \mathbf{a}_2(t-\tau) & t \geq \tau. \end{cases}$$

Similarly, given $a \in \mathcal{A}$ and $\mathbf{a} \in \mathbb{A}$, we let $[a, \mathbf{a}]$ be the element of $\mathbb{A}$ for which

$$[a, \mathbf{a}](t) = \begin{cases} a & t = 0, \\ \mathbf{a}(t-1) & t \geq 1. \end{cases}$$

Additionally, given $\mathbf{a} \in \mathbb{A}$ and $\tau \in \mathbb{N}$, we let $\mathbf{a}|_\tau$ be the element of $\mathbb{A}$ for which

$$\mathbf{a}|_\tau(t) = \mathbf{a}(t+\tau) \quad \forall t \in \mathbb{N}.$$

The $[\cdot,\cdot]_\tau$ operation corresponds to concatenating two action sequences (using only the 0^{th} to $(\tau-1)^{\text{st}}$ elements of the first sequence), the $[\cdot,\cdot]$ operation corresponds to prepending an action to an action sequence, and the $\cdot|_\tau$ operation corresponds to removing the 0^{th} to $(\tau-1)^{\text{st}}$ elements of an action sequence.

We let Π be the set of policies $\pi : \mathcal{S} \to \mathcal{A}$. Given $s \in \mathcal{S}$ and $\pi \in \Pi$, we let $\xi_s^\pi : \mathbb{N} \to \mathcal{S}$ be the solution of the evolution equation

$$\xi_s^\pi(t+1) = f\left(\xi_s^\pi(t), \pi\left(\xi_s^\pi(t)\right)\right)$$

for which $\xi_s^\pi(0) = s$. In other words, $\xi_s^\pi(\cdot)$ is the state trajectory over time when the agent begins at state s and follows policy π.

We will also "overload" this trajectory notation for signals rather than policies: given $\mathbf{a} \in \mathbb{A}$, we let $\xi_s^{\mathbf{a}} : \mathbb{N} \to \mathcal{S}$ be the solution of the evolution equation

$$\xi_s^{\mathbf{a}}(t+1) = f\left(\xi_s^{\mathbf{a}}(t), \mathbf{a}(t)\right)$$

for which $\xi_s^{\mathbf{a}}(0) = s$. In other words, $\xi_s^{\mathbf{a}}(\cdot)$ is the state trajectory over time when the agent begins at state s and follows action sequence $\mathbf{a}$.

A Proof of RAA Main Theorem

We first define the value functions, $V_{\text{A}}^*, \tilde{V}_{\text{RA}}^*, V_{\text{RAA}}^* : \mathcal{S} \to \mathbb{R}$ by

$$\begin{aligned} V_{\text{A}}^*(s) &= \max_{\pi \in \Pi} \min_{\tau \in \mathbb{N}} q\left(\xi_s^\pi(\tau)\right), \\ \tilde{V}_{\text{RA}}^*(s) &= \max_{\pi \in \Pi} \max_{\tau \in \mathbb{N}} \min\left\{ r_{\text{RAA}}\left(\xi_s^\pi(\tau)\right), \min_{\kappa \leq \tau} q\left(\xi_s^\pi(\kappa)\right) \right\}, \\ V_{\text{RAA}}^*(s) &= \max_{\pi \in \Pi} \min\left\{ \max_{\tau \in \mathbb{N}} r\left(\xi_s^\pi(\tau)\right), \min_{\kappa \in \mathbb{N}} q\left(\xi_s^\pi(\kappa)\right) \right\}, \end{aligned}$$

where r_{RAA} is as in Theorem 1.

We next define the value functions, $v_{\text{A}}^*, \tilde{v}_{\text{RA}}^*, v_{\text{RAA}}^* : \mathcal{S} \to \mathbb{R}$, which maximize over action sequences rather than policies:

$$\begin{aligned} v_{\text{A}}^*(s) &= \max_{\mathbf{a} \in \mathbb{A}} \min_{\tau \in \mathbb{N}} q\left(\xi_s^{\mathbf{a}}(\tau)\right), \\ \tilde{v}_{\text{RA}}^*(s) &= \max_{\mathbf{a} \in \mathbb{A}} \max_{\tau \in \mathbb{N}} \min\left\{ r_{\text{RAA}}\left(\xi_s^{\mathbf{a}}(\tau)\right), \min_{\kappa \leq \tau} q\left(\xi_s^{\mathbf{a}}(\kappa)\right) \right\}, \\ v_{\text{RAA}}^*(s) &= \max_{\mathbf{a} \in \mathbb{A}} \min\left\{ \max_{\tau \in \mathbb{N}} r\left(\xi_s^{\mathbf{a}}(\tau)\right), \min_{\kappa \in \mathbb{N}} q\left(\xi_s^{\mathbf{a}}(\kappa)\right) \right\}, \end{aligned}$$

Observe that for each $s \in \mathcal{S}$,

$$v_{\text{A}}^*(s) \geq V_{\text{A}}^*(s), \quad \tilde{v}_{\text{RA}}^*(s) \geq \tilde{V}_{\text{RA}}^*(s), \quad v_{\text{RAA}}^*(s) \geq V_{\text{RAA}}^*(s).$$

We now prove a series of lemmas that will be useful in the proof of the main theorem.

Lemma 1. *There is a $\pi \in \Pi$ such that*

$$v_{\mathrm{A}}^*(s) = \min_{\tau \in \mathbb{N}} q\left(\xi_s^\pi(\tau)\right)$$

for all $s \in \mathcal{S}$.

Proof. Choose $\pi \in \Pi$ such that

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} v_{\mathrm{A}}^*\left(f(s,a)\right) \quad \forall s \in \mathcal{S}.$$

Fix $s \in \mathcal{S}$. Note that for each $\tau \in \mathbb{N}$,

$$\begin{aligned}
v_{\mathrm{A}}^*\left(\xi_s^\pi(\tau+1)\right) &= v_{\mathrm{A}}^*\left(f\left(\xi_s^\pi(\tau), \pi\left(\xi_s^\pi(\tau)\right)\right)\right) \\
&= \max_{a \in \mathcal{A}} v_{\mathrm{A}}^*\left(f\left(\xi_s^\pi(\tau), a\right)\right) \\
&= \max_{a \in \mathcal{A}} \max_{\mathbf{a} \in \mathbb{A}} \min_{\kappa \in \mathbb{N}} q\left(\xi^{\mathbf{a}}_{f\left(\xi_s^\pi(\tau), a\right)}(\kappa)\right) \\
&= \max_{a \in \mathcal{A}} \max_{\mathbf{a} \in \mathbb{A}} \min_{\kappa \in \mathbb{N}} q\left(\xi^{[a,\mathbf{a}]}_{\xi_s^\pi(\tau)}(\kappa+1)\right) \\
&= \max_{\mathbf{a} \in \mathbb{A}} \min_{\kappa \in \mathbb{N}} q\left(\xi^{\mathbf{a}}_{\xi_s^\pi(\tau)}(\kappa+1)\right) \\
&\geq \max_{\mathbf{a} \in \mathbb{A}} \min_{\kappa \in \mathbb{N}} q\left(\xi^{\mathbf{a}}_{\xi_s^\pi(\tau)}(\kappa)\right) \\
&\geq v_{\mathrm{A}}^*\left(\xi_s^\pi(\tau)\right).
\end{aligned}$$

It follows by induction that $v_{\mathrm{A}}^*\left(\xi_s^\pi(\tau)\right) \geq v_{\mathrm{A}}^*\left(\xi_s^\pi(0)\right)$ for all $\tau \in \mathbb{N}$, so that

$$v_{\mathrm{A}}^*(s) \geq \min_{\tau \in \mathbb{N}} q\left(\xi_s^\pi(\tau)\right) \geq \min_{\tau \in \mathbb{N}} v_{\mathrm{A}}^*\left(\xi_s^\pi(\tau)\right) = v_{\mathrm{A}}^*\left(\xi_s^\pi(0)\right) = v_{\mathrm{A}}^*(s).$$

□

Corollary 3. *For all $s \in \mathcal{S}$, we have $V_{\mathrm{A}}^*(s) = v_{\mathrm{A}}^*(s)$.*

Lemma 2. *There is a $\pi \in \Pi$ such that*

$$\tilde{v}_{\mathrm{RA}}^*(s) = \max_{\tau \in \mathbb{N}} \min\left\{ r_{\mathrm{RAA}}\left(\xi_s^\pi(\tau)\right), \min_{\kappa \leq \tau} q\left(\xi_s^\pi(\kappa)\right) \right\}$$

for all $s \in \mathcal{S}$.

Proof. First, let us note that in this proof we will use the standard conventions that

$$\max \varnothing = -\infty \quad \text{and} \quad \min \varnothing = +\infty.$$

We next introduce some notation. First, for convenience, we set $v^* = \tilde{v}_{\mathrm{RA}}^*$ and $V^* = \tilde{V}_{\mathrm{RA}}^*$. Given $s \in \mathcal{S}$ and $\mathbf{a} \in \mathbb{A}$, we write

$$v^{\mathbf{a}}(s) = \max_{\tau \in \mathbb{N}} \min\left\{ r_{\mathrm{RAA}}\left(\xi_s^{\mathbf{a}}(\tau)\right), \min_{\kappa \leq \tau} q\left(\xi_s^{\mathbf{a}}(\kappa)\right) \right\}.$$

Similarly, given $s \in \mathcal{S}$ and $\pi \in \Pi$, we write

$$V^\pi(s) = \max_{\tau \in \mathbb{N}} \min\left\{ r_{\mathrm{RAA}}\left(\xi_s^\pi(\tau)\right), \min_{\kappa \leq \tau} q\left(\xi_s^\pi(\kappa)\right) \right\}.$$

Then

$$V^*(s) = \max_{\pi \in \Pi} \max_{\tau \in \mathbb{N}} \min\left\{ r_{\mathrm{RAA}}\left(\xi_s^\pi(\tau)\right), \min_{\kappa \leq \tau} q\left(\xi_s^\pi(\kappa)\right) \right\} = \max_{\pi \in \Pi} V^\pi(s),$$

and

$$v^*(s) = \max_{\mathbf{a} \in \mathbb{A}} \max_{\tau \in \mathbb{N}} \min\left\{ r_{\mathrm{RAA}}\left(\xi_s^{\mathbf{a}}(\tau)\right), \min_{\kappa \leq \tau} q\left(\xi_s^{\mathbf{a}}(\kappa)\right) \right\} = \max_{\mathbf{a} \in \mathbb{A}} v^{\mathbf{a}}(s).$$

It is immediate that $v^*(s) \geq V^*(s)$ for each $s \in \mathcal{S}$, so it suffices to show the reverse inequality. Toward this end, it suffices to show that there is a $\pi \in \Pi$ for which $V^\pi(s) = v^*(s)$ for each $s \in \mathcal{S}$. Indeed, in this case, $V^*(s) \geq V^\pi(s) = v^*(s)$.

We now construct the desired policy π. Let $\alpha_0 = +\infty$, $S_0 = \varnothing$, and $v_0^* : \mathcal{S} \to \mathbb{R} \cup \{-\infty\}, s \mapsto -\infty$. We recursively define $\alpha_t \in \mathbb{R}$, $S_t \subseteq \mathcal{S}$, and $v_t^* : \mathcal{S} \to \mathbb{R} \cup \{-\infty\}$ for $t = 1, 2, \ldots$ by

$$\alpha_{t+1} = \max_{s \in \mathcal{S} \setminus S_t} \min \left\{ \max \left\{ r_{\text{RAA}}(s), \max_{a \in \mathcal{A}} v_t^* \left(f(s,a) \right) \right\}, q(s) \right\}, \tag{3}$$

$$S_{t+1} = S_t \cup \left\{ s \in \mathcal{S} \setminus S_t \;\middle|\; \min \left\{ \max \left\{ r_{\text{RAA}}(s), \max_{a \in \mathcal{A}} v_t^* \left(f(s,a) \right) \right\}, q(s) \right\} = \alpha_{t+1} \right\}, \tag{4}$$

$$v_{t+1}^*(s) = \begin{cases} v_t^*(s) & s \in S_t, \\ \alpha_{t+1} & s \in S_{t+1} \setminus S_t, \\ -\infty & s \in \mathcal{S} \setminus S_{t+1}. \end{cases} \tag{5}$$

From (4) it follows that

$$S_0 \subseteq S_1 \subseteq S_2 \subseteq \ldots, \tag{6}$$

which together with (3) shows that

$$\alpha_0 \geq \alpha_1 \geq \alpha_2 \geq \ldots. \tag{7}$$

Also, whenever $\mathcal{S} \setminus S_t$ is non-empty, the set being appended to S_t in (4) is non-empty so

$$\bigcup_{t=0}^{\infty} S_t = \mathcal{S}. \tag{8}$$

For each $s \in \mathcal{S}$, let $\sigma(s)$ be the smallest $t \in \mathbb{N}$ for which $s \in S_t$. We choose the policy $\pi \in \Pi$ of interest by insisting

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} v_{\sigma(s)-1}^*(f(s,a)) \quad \forall s \in \mathcal{S}. \tag{9}$$

In the remainder of the proof, we show that $V^\pi(s) = v^*(s)$ for each $s \in \mathcal{S}$ by induction. Let $n \in \mathbb{N}$ and suppose the following induction assumptions hold:

$$V^\pi(s) = v^*(s) = v_n^*(s) \geq \alpha_n \quad \forall s \in S_n, \tag{10}$$

$$v^*(s') \leq \alpha_n \quad \forall s' \in \mathcal{S} \setminus S_n. \tag{11}$$

Note that the above hold trivially when $n = 0$ since $S_0 = \varnothing$ and $\alpha_0 = +\infty$. Fix some particular $y \in S_{n+1}$ and some $z \in \mathcal{S} \setminus S_{n+1}$. We must show that

$$V^\pi(y) = v^*(y) = v_{n+1}^*(y) \geq \alpha_{n+1}, \tag{12}$$

$$v^*(z) \leq \alpha_{n+1}. \tag{13}$$

In this case, induction then shows that $V^\pi(s) = v^*(s)$ for all $s \in \cup_{n=0}^{\infty} S_t$. Since this union is equal to $\mathcal{S}$ by (8), the desired result then follows.

To show (12)-(13), we first demonstrate the following three claims.

1. Let $x \in \mathcal{S}$ and $w \in \mathcal{A}$ be such that $f(x,w) \in S_n$ and $q(x) \geq \alpha_{n+1}$. We claim $x \in S_{n+1}$.

   We can assume $x \notin S_n$, for otherwise the claim follows immediately from (6). Since $f(x,w) \in S_n$, we have $v_n^* \left(f(x,w) \right) \geq \alpha_n$ by (10). Thus

   $$\begin{aligned} \alpha_{n+1} &\geq \min \left\{ \max \left\{ r_{\text{RAA}}(x), \max_{a \in \mathcal{A}} v_n^* \left(f(x,a) \right) \right\}, q(x) \right\} \\ &\geq \min \left\{ \max\{ r_{\text{RAA}}(x), \alpha_n \}, \alpha_{n+1} \right\} \\ &= \alpha_{n+1}, \end{aligned}$$

   where the first inequality follows from (3), and the equality follows from (7). Thus

   $$\alpha_{n+1} = \min \left\{ \max \left\{ r_{\text{RAA}}(x), \max_{a \in \mathcal{A}} v_n^* \left(f(x,a) \right) \right\}, q(x) \right\},$$

   so the claim follows from (4).

2. Let $x \in S_{n+1} \setminus S_n$ and $w \in \mathcal{A}$ be such that $f(x,w) \in S_n$. We claim that
$$V^\pi(x) = v^*(x) = \alpha_{n+1}. \tag{14}$$
To show this claim, we will make use of the dynamic programming principle
$$v^{\mathbf{a}}(s) = \min\left\{\max\left\{r_{\text{RAA}}(s), v^{\mathbf{a}|_1}\left(f(s,\mathbf{a}(0))\right)\right\}, q(s)\right\}, \quad \forall s \in \mathcal{S}, \mathbf{a} \in \mathbb{A},$$
from which it follows that
$$V^\pi(s) = \min\left\{\max\left\{r_{\text{RAA}}(s), V^\pi\left(f(s,\pi(s))\right)\right\}, q(s)\right\}, \quad \forall s \in \mathcal{S}, \tag{15}$$
and
$$v^*(s) = \min\left\{\max\left\{r_{\text{RAA}}(s), \max_{a\in\mathcal{A}} v^*\left(f(s,a)\right)\right\}, q(s)\right\}, \quad \forall s \in \mathcal{S}. \tag{16}$$
Since $x \in S_{n+1} \setminus S_n$, then $\sigma(x) = n+1$ by definition of σ, so $\pi(x) \in \arg\max_{a\in\mathcal{A}} v_n^*(f(x,a))$ by (9). Thus
$$v_n^*\left(f(x,\pi(x))\right) = \max_{a\in\mathcal{A}} v_n^*\left(f(x,a)\right). \tag{17}$$
But then
$$v_n^*\left(f(x,\pi(x))\right) \geq v_n^*\left(f(x,w)\right) \geq \alpha_n \geq \alpha_{n+1} > -\infty,$$
where the second inequality comes from (10), the third comes from (7), and the final inequality comes from (3) ($\mathcal{S} \setminus S_n$ is non-empty because $x \in \mathcal{S} \setminus S_n$). Thus $f(x,\pi(x)) \in S_n$ by (5). It then follows from (10) that
$$V^\pi\left(f(x,\pi(x))\right) = v^*\left(f(x,\pi(x))\right) = v_n^*\left(f(x,\pi(x))\right). \tag{18}$$
Now, observe that for all $s \in S_n$ and $s' \in \mathcal{S} \setminus S_n$,
$$v^*(s) = v_n^*(s) \geq \alpha_n \geq v^*(s') \geq -\infty = v_n^*(s'), \tag{19}$$
where the first equality and inequality are from (10), the second inequality is from (11), and the final equality is from (5). Moreover, $f(x,a) \in S_n$ for at least one a (in particular $a = w$). Letting $\mathcal{A}' = \{a \in \mathcal{A} \mid f(x,a) \in S_n\}$, it follows from (19) that
$$\max_{a\in\mathcal{A}} v^*\left(f(x,a)\right) = \max_{a\in\mathcal{A}'} v^*\left(f(x,a)\right) = \max_{a\in\mathcal{A}'} v_n^*\left(f(x,a)\right) = \max_{a\in\mathcal{A}} v_n^*\left(f(x,a)\right). \tag{20}$$
From (17)-(20) we have
$$V^\pi\left(f(x,\pi(x))\right) = \max_{a\in\mathcal{A}} v^*\left(f(x,a)\right) = \max_{a\in\mathcal{A}} v_n^*\left(f(x,a)\right). \tag{21}$$
Now observe that
$$\begin{aligned}
V^\pi(x) &= \min\left\{\max\left\{r_{\text{RAA}}(x), V^\pi\left(f(x,\pi(x))\right)\right\}, q(x)\right\}, \\
v^*(x) &= \min\left\{\max\left\{r_{\text{RAA}}(x), \max_{a\in\mathcal{A}} v^*\left(f(x,a)\right)\right\}, q(x)\right\}, \\
\alpha_{n+1} &= \min\left\{\max\left\{r_{\text{RAA}}(x), \max_{a\in\mathcal{A}} v_n^*\left(f(x,a)\right)\right\}, q(x)\right\},
\end{aligned}$$
where the first equation is from (15), the second is from (16), and the third is from (4). But then (14) follows from the above equations together with (21).

3. Let $x \in \mathcal{S} \setminus S_n$. We claim that $v^*(x) \leq \alpha_{n+1}$. Suppose otherwise. Then we can choose $\mathbf{a} \in \mathbb{A}$ and $\tau \in \mathbb{N}$ such that
$$\min\left\{r_{\text{RAA}}\left(\xi_x^{\mathbf{a}}(\tau)\right), \min_{\kappa\leq\tau} q\left(\xi_x^{\mathbf{a}}(\kappa)\right)\right\} > \alpha_{n+1}. \tag{22}$$
It follows that $\xi_x^{\mathbf{a}}(\tau) \in S_n$, for otherwise
$$\alpha_{n+1} \geq \min\left\{r_{\text{RAA}}(\xi_x^{\mathbf{a}}(\tau)), q(\xi_x^{\mathbf{a}}(\tau))\right\}$$

by (3), creating a contradiction.

So $x \notin S_n$ and $\xi_x^{\mathbf{a}}(\tau) \in S_n$, indicating that there is some $\theta \in \{0, \dots, \tau - 1\}$ such that $\xi_x^{\mathbf{a}}(\theta) \notin S_n$ and $f\left(\xi_x^{\mathbf{a}}(\theta), \mathbf{a}(\theta)\right) = \xi_x^{\mathbf{a}}(\theta + 1) \in S_n$. Moreover, $q\left(\xi_x^{\mathbf{a}}(\theta)\right) > \alpha_{n+1}$ by (22). It follows from claim 1 that $\xi_x^{\mathbf{a}}(\theta) \in S_{n+1}$.

But then it follows from claim 2 that $v^*\left(\xi_x^{\mathbf{a}}(\theta)\right) = \alpha_{n+1}$. However,

$$\begin{aligned} v^*\left(\xi_x^{\mathbf{a}}(\theta)\right) &\geq \min\left\{ r_{\text{RAA}}\left(\xi^{\mathbf{a}}_{\xi_x^{\mathbf{a}}(\theta)}(\tau - \theta)\right), \min_{\kappa \leq \tau - \theta} q\left(\xi^{\mathbf{a}}_{\xi_x^{\mathbf{a}}(\theta)}(\kappa)\right)\right\} \\ &= \min\left\{ r_{\text{RAA}}\left(\xi_x^{\mathbf{a}}(\tau - \theta + \theta)\right), \min_{\kappa \leq \tau - \theta} q\left(\xi_x^{\mathbf{a}}(\kappa + \theta)\right)\right\} \\ &= \min\left\{ r_{\text{RAA}}\left(\xi_x^{\mathbf{a}}(\tau)\right), \min_{\kappa \in \{\theta, \theta+1, \dots, \tau\}} q\left(\xi_x^{\mathbf{a}}(\kappa)\right)\right\} \\ &> \alpha_{n+1}, \end{aligned}$$

giving the desired contradiction.

Having established these claims, we return to proving (12) and (13) hold. In fact, (13) follows immediately from claim 3, so we actually only need to show (12).

If $y \in S_n$, then from (5) and (10), we have that $V^\pi(y) = v^*(y) = v_n^*(y) = v_{n+1}^*(y)$, and from (7) and (10), we also have that $v_n^*(y) \geq \alpha_n \geq \alpha_{n+1}$. Together these establish (12) when $y \in S_n$.

So suppose $y \in S_{n+1} \setminus S_n$. First, observe that $v_{n+1}^*(y) = \alpha_{n+1}$ by (5). There are now two possibilities. If there is some $a \in \mathcal{A}$ for which $f(y, a) \in S_n$, then (12) follows from claim 2. If instead, $f(y, a) \notin S_n$ for each $a \in \mathcal{A}$, then $\max_{a \in \mathcal{A}} v_n^*\left(f(y, a)\right) = -\infty$ by (5) (or if $n = 0$ by definition of v_0^*). Thus $\alpha_{n+1} = \min\left\{r_{\text{RAA}}(y), q(y)\right\}$ by (4), so

$$v^*(y) \geq V^\pi(y) \geq \min\left\{r_{\text{RAA}}(y), q(y)\right\} = \alpha_{n+1} \geq v^*(y),$$

where the final inequality follows from claim 3. This completes the proof. □

Corollary 4. *For all $s \in \mathcal{S}$, we have $\tilde{V}_{\text{RA}}^*(s) = \tilde{v}_{\text{RA}}^*(s)$.*

Lemma 3. *Let $F : \mathbb{A} \times \mathbb{N} \to \mathbb{R}$. Then*

$$\sup_{\mathbf{a} \in \mathbb{A}} \sup_{\tau \in \mathbb{N}} \sup_{\mathbf{a}' \in \mathbb{A}'} F\left([\mathbf{a}, \mathbf{a}']_\tau, \tau\right) = \sup_{\mathbf{a} \in \mathbb{A}} \sup_{\tau \in \mathbb{N}} F\left(\mathbf{a}, \tau\right). \tag{23}$$

Proof. We proceed by showing both inequalities corresponding to (23) hold.

- ($\geq$) Given any $\mathbf{a} \in \mathbb{A}$ and $\tau \in \mathbb{N}$, we have $\sup_{\mathbf{a}' \in \mathbb{A}'} F\left([\mathbf{a}, \mathbf{a}']_\tau, \tau\right) \geq F\left(\mathbf{a}, \tau\right)$. Taking the suprema over $\mathbf{a} \in \mathbb{A}$ and $\tau \in \mathbb{N}$ on both sides of this inequality gives the desired result.
- ($\leq$) Given any $\mathbf{a} \in \mathbb{A}$ and $\tau \in \mathbb{N}$, we have

  $$\sup_{\mathbf{a}' \in \mathbb{A}'} F\left([\mathbf{a}, \mathbf{a}']_\tau, \tau\right) \leq \sup_{\mathbf{a}'' \in \mathbb{A}} F\left(\mathbf{a}'', \tau\right),$$

  so that the result follows from taking the suprema over $\mathbf{a} \in \mathbb{A}$ and $\tau \in \mathbb{N}$ on both sides of this inequality.

□

Lemma 4. *For each $s \in \mathcal{S}$,*

$$v_{\text{RAA}}^*(s) = \tilde{v}_{\text{RA}}^*(s).$$

Proof. For each $s \in \mathcal{S}$, we have

$$\tilde{v}^*_{\text{RA}}(s) = \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \min \left\{ r_{\text{RAA}}\left(\xi^{\mathbf{a}}_s(\tau)\right), \min_{\kappa\le\tau} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\} \tag{24}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \min \left\{ r\left(\xi^{\mathbf{a}}_s(\tau)\right), v^*_{\text{A}}\left(\xi^{\mathbf{a}}_s(\tau)\right), \min_{\kappa\le\tau} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\} \tag{25}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \min \left\{ r\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\mathbf{a}'\in\mathbb{A}} \min_{\kappa'\in\mathbb{N}} q\left(\xi^{\mathbf{a}'}_{\xi^{\mathbf{a}}_s(\tau)}(\kappa')\right), \min_{\kappa\le\tau} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \min \left\{ r\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\mathbf{a}'\in\mathbb{A}} \min_{\kappa'\in\mathbb{N}} q\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\kappa')\right), \min_{\kappa\le\tau} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max_{\mathbf{a}'\in\mathbb{A}} \min \left\{ r\left(\xi^{\mathbf{a}}_s(\tau)\right), \min_{\kappa'\in\mathbb{N}} q\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\kappa')\right), \min_{\kappa\le\tau} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max_{\mathbf{a}'\in\mathbb{A}} \min \left\{ r\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau)\right), \min_{\kappa'\in\mathbb{N}} q\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\kappa')\right), \min_{\kappa\le\tau} q\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\kappa)\right) \right\} \tag{26}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \min \left\{ r\left(\xi^{\mathbf{a}}_s(\tau)\right), \min_{\kappa'\in\mathbb{N}} q\left(\xi^{\mathbf{a}}_s(\tau+\kappa')\right), \min_{\kappa\le\tau} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\} \tag{27}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \min \left\{ r\left(\xi^{\mathbf{a}}_s(\tau)\right), \min_{\kappa\in\mathbb{N}} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\}$$

$$= \max_{\mathbf{a}\in\mathbb{A}} \min \left\{ \max_{\tau\in\mathbb{N}} r\left(\xi^{\mathbf{a}}_s(\tau)\right), \min_{\kappa\in\mathbb{N}} q\left(\xi^{\mathbf{a}}_s(\kappa)\right) \right\}$$

$$= v^*_{\text{RAA}}(s),$$

where the equality between (24) and (25) follows from Corollary 3, and where the equality between (26) and (27) follows from Lemma 3. □

Before the next lemma, we need to introduce two last pieces of notation. First, we let $\overline{\Pi}$ be the set of augmented policies $\bar{\pi} : \mathcal{S} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{A}$, where

$$\mathcal{Y} = \{r(s) \mid s \in \mathcal{S}\} \quad \text{and} \quad \mathcal{Z} = \{q(s) \mid s \in \mathcal{S}\}.$$

Next, given $s \in \mathcal{S}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, and $\bar{\pi} \in \overline{\Pi}$, we let $\bar{\xi}^{\bar{\pi}}_s : \mathbb{N} \to \mathcal{S}$, $\bar{\eta}^{\bar{\pi}}_s : \mathbb{N} \to \mathcal{Y}$, and $\bar{\zeta}^{\bar{\pi}}_s : \mathbb{N} \to \mathcal{Z}$, be the solution of the evolution

$$\begin{aligned}
\bar{\xi}^{\bar{\pi}}_s(t+1) &= f\left(\bar{\xi}^{\bar{\pi}}_s(t), \bar{\pi}\left(\bar{\xi}^{\bar{\pi}}_s(t), \bar{\eta}^{\bar{\pi}}_s(t), \bar{\zeta}^{\bar{\pi}}_s(t)\right)\right), \\
\bar{\eta}^{\bar{\pi}}_s(t+1) &= \max\left\{r\left(\bar{\xi}^{\bar{\pi}}_s(t+1)\right), \bar{\eta}^{\bar{\pi}}_s(t)\right\}, \\
\bar{\zeta}^{\bar{\pi}}_s(t+1) &= \min\left\{q\left(\bar{\xi}^{\bar{\pi}}_s(t+1)\right), \bar{\zeta}^{\bar{\pi}}_s(t)\right\},
\end{aligned}$$

for which $\bar{\xi}^{\bar{\pi}}_s(0) = s$, $\bar{\eta}^{\bar{\pi}}_s(0) = r(s)$, and $\bar{\zeta}^{\bar{\pi}}_s(0) = q(s)$.

Lemma 5. *There is a $\bar{\pi} \in \overline{\Pi}$ such that*

$$v^*_{\text{RAA}}(s) = \min\left\{\max_{\tau\in\mathbb{N}} r\left(\bar{\xi}^{\bar{\pi}}_s(\tau)\right), \min_{\tau\in\mathbb{N}} q\left(\bar{\xi}^{\bar{\pi}}_s(\tau)\right)\right\} \tag{28}$$

for all $s \in \mathcal{S}$.

Proof. By Lemmas 1 and 2 together with Corollary 3, we can choose $\pi, \theta \in \Pi$ such that

$$\begin{aligned}
\tilde{v}^*_{\text{RA}}(s) &= \max_{\tau\in\mathbb{N}} \min\left\{r\left(\xi^{\pi}_s(\tau)\right), v^*_{\text{A}}\left(\xi^{\pi}_s(\tau)\right), \min_{\kappa\le\tau} q\left(\xi^{\pi}_s(\kappa)\right)\right\} \quad \forall s \in \mathcal{S}, \\
v^*_{\text{A}}(s) &= \min_{\tau\in\mathbb{N}} q\left(\xi^{\theta}_s(\tau)\right) \quad \forall s \in \mathcal{S}.
\end{aligned}$$

We introduce some useful notation we will use throughout the rest of the proof. For each $s \in \mathcal{S}$, let $[s]^+ = f(s, \pi(s))$, $[y]^+_s = \max\{y, r([s]^+)\}$, $[z]^+_s = \min\{z, q([s]^+)\}$.

We define an augmented policy $\bar{\pi} \in \overline{\Pi}$ by

$$\bar{\pi}(s,y,z) = \begin{cases} \pi(s) & \min\{[y]_s^+, [z]_s^+, v_{\mathrm{A}}^*([s]^+)\} \geq \min\{y, z, v_{\mathrm{A}}^*(s)\}, \\ \theta(s) & \text{otherwise.} \end{cases}$$

Now fix some $s \in \mathcal{S}$. For all $t \in \mathbb{N}$, set $\bar{x}_t = \bar{\xi}_s^{\bar{\pi}}(t)$, $\bar{y}_t = \bar{\eta}_s^{\bar{\pi}}(t) = \max_{\tau \leq t} r(\bar{x}_\tau)$, and $\bar{z}_t = \bar{\zeta}_s^{\bar{\pi}}(t) = \min_{\tau \leq t} q(\bar{x}_\tau)$, and also set $x_t^\circ = \xi_s^\pi(t)$, $y_t^\circ = \max_{\tau \leq t} r(x_\tau^\circ)$, and $z_t^\circ = \min_{\tau \leq t} q(x_\tau^\circ)$.

First, assume that t is such that $\min\{[\bar{y}_t]_{\bar{x}_t}^+, [\bar{z}_t]_{\bar{x}_t}^+, v_{\mathrm{A}}^*([\bar{x}_t]^+)\} < \min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\}$. In this case, $\bar{\pi}(\bar{x}_t, \bar{y}_t, \bar{z}_t) = \theta(\bar{x}_t)$, so that

$$\min\{\bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\} = \min\{\bar{z}_{t+1}, v_{\mathrm{A}}^*(\bar{x}_{t+1})\}$$

by our choice of θ. Since $\bar{y}_t$ is non-decreasing in t, thus have

$$\min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\} \leq \min\{\bar{y}_{t+1}, \bar{z}_{t+1}, v_{\mathrm{A}}^*(\bar{x}_{t+1})\}.$$

Next, assume that t is such that $\min\{[\bar{y}_t]_{\bar{x}_t}^+, [\bar{z}_t]_{\bar{x}_t}^+, v_{\mathrm{A}}^*([\bar{x}_t]^+)\} \geq \min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\}$. In this case, we have that $\bar{\pi}(\bar{x}_t, \bar{y}_t, \bar{z}_t) = \pi(\bar{x}_t)$, so

$$\min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\} \leq \min\{[\bar{y}_t]_{\bar{x}_t}^+, [\bar{z}_t]_{\bar{x}_t}^+, v_{\mathrm{A}}^*([\bar{x}_t]^+)\} = \min\{\bar{y}_{t+1}, \bar{z}_{t+1}, v_{\mathrm{A}}^*(\bar{x}_{t+1})\}.$$

It thus follows from these two cases that $\min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\}$ is non-decreasing in t. Let

$$T = \min\left\{t \in \mathbb{N} \mid \min\{[\bar{y}_t]_{\bar{x}_t}^+, [\bar{z}_t]_{\bar{x}_t}^+, v_{\mathrm{A}}^*([\bar{x}_t]^+)\} < \min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\}\right\}.$$

There are again two cases:

($T < \infty$) In this case, $\bar{\pi}(\bar{x}_t, \bar{y}_t, \bar{z}_t) = \pi(\bar{x}_t)$ for $t < T$. Then $\bar{x}_t = x_t^\circ$, $\bar{y}_t = y_t^\circ$, and $\bar{z}_t = z_t^\circ$ for all $t \leq T$. It follows that $[\bar{x}_t]^+ = x_{t+1}^\circ$, $[\bar{y}_t]_{\bar{x}_t}^+ = y_{t+1}^\circ$, and $[\bar{z}_t]_{\bar{x}_t}^+ = z_{t+1}^\circ$ for all $t \leq T$. Thus by definition of T,

$$\min\left\{y_{t+1}^\circ, z_{t+1}^\circ, v_{\mathrm{A}}^*\left(x_{t+1}^\circ\right)\right\} \geq \min\left\{y_t^\circ, z_t^\circ, v_{\mathrm{A}}^*\left(x_t^\circ\right)\right\} \quad \forall t < T.$$

and

$$\min\left\{y_{T+1}^\circ, z_{T+1}^\circ, v_{\mathrm{A}}^*\left(x_{T+1}^\circ\right)\right\} < \min\left\{y_T^\circ, z_T^\circ, v_{\mathrm{A}}^*\left(x_T^\circ\right)\right\}.$$

But since y_t° is non-decreasing and $\min\{z_t^\circ, v_{\mathrm{A}}^*(x_t^\circ)\}$ is non-increasing in t, it follows that $\min\{y_t^\circ, z_t^\circ, v_{\mathrm{A}}^*(x_t^\circ)\}$ must achieve its maximal value at the smallest t for which it strictly decreases from t to $t+1$, i.e.

$$\begin{aligned} \min\left\{\bar{y}_T, \bar{z}_T, v_{\mathrm{A}}^*\left(\bar{x}_T\right)\right\} &= \min\left\{y_T^\circ, z_T^\circ, v_{\mathrm{A}}^*\left(x_T^\circ\right)\right\} \\ &= \max_{t \in \mathbb{N}} \min\left\{y_t^\circ, z_t^\circ, v_{\mathrm{A}}^*\left(x_t^\circ\right)\right\} \\ &\geq \max_{t \in \mathbb{N}} \min\left\{r\left(x_t^\circ\right), z_t^\circ, v_{\mathrm{A}}^*\left(x_t^\circ\right)\right\} \\ &= \tilde{v}_{\mathrm{RA}}^*(s). \end{aligned}$$

where the final equality follows from our choice of π. Since $\min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\}$ is non-decreasing in t, then

$$\min\{\bar{y}_t, \bar{z}_t\} \geq \min\{\bar{y}_t, \bar{z}_t, v_{\mathrm{A}}^*(\bar{x}_t)\} \geq \min\{\bar{y}_T, \bar{z}_T, v_{\mathrm{A}}^*(\bar{x}_T)\} = \tilde{v}_{\mathrm{RA}}^*(s) \quad \forall t \geq T.$$

Thus

$$v_{\mathrm{RAA}}^*(s) \geq \min\left\{\max_{t \in \mathbb{N}} r\left(\bar{x}_t\right), \min_{t \in \mathbb{N}} q\left(\bar{x}_t\right)\right\} = \lim_{t \to \infty} \min\{\bar{y}_t, \bar{z}_t\} \geq \tilde{v}_{\mathrm{RA}}^*(s) = v_{\mathrm{RAA}}^*(s),$$

where the final equality follows from Lemma (4). Thus the proof is complete in this case.

($T = \infty$) In this case, $\bar{\pi}(\bar{x}_t, \bar{y}_t, \bar{z}_t) = \pi(\bar{x}_t)$ for all $t \in \mathbb{N}$. Then $\bar{x}_t = x_t^\circ$, $\bar{y}_t = y_t^\circ$, and $\bar{z}_t = z_t^\circ$ for all $t \in \mathbb{N}$. Also $[\bar{x}_t]^+ = x_{t+1}^\circ$, $[\bar{y}_t]_{\bar{x}_t}^+ = y_{t+1}^\circ$, and $[\bar{z}_t]_{\bar{x}_t}^+ = z_{t+1}^\circ$ for all $t \in \mathbb{N}$. Thus by definition of T,

$$\min\left\{y_{t+1}^\circ, z_{t+1}^\circ, v_{\mathrm{A}}^*\left(x_{t+1}^\circ\right)\right\} \geq \min\left\{y_t^\circ, z_t^\circ, v_{\mathrm{A}}^*\left(x_t^\circ\right)\right\} \quad \forall t \in \mathbb{N}.$$

Let $T' \in \arg\max_{t\in\mathbb{N}} \min\{y_t^\circ, z_t^\circ, v_{\mathrm{A}}^*(x_t^\circ)\}$. Then

$$\begin{aligned}\min\{\bar{y}_{T'}, \bar{z}_{T'}, v_{\mathrm{A}}^*(\bar{x}_{T'})\} &= \min\{y_{T'}^\circ, z_{T'}^\circ, v_{\mathrm{A}}^*(x_{T'}^\circ)\} \\ &= \max_{t\in\mathbb{N}} \min\{y_t^\circ, z_t^\circ, v_{\mathrm{A}}^*(x_t^\circ)\} \\ &\geq \max_{t\in\mathbb{N}} \min\{r(x_t^\circ), z_t^\circ, v_{\mathrm{A}}^*(x_t^\circ)\} \\ &= \tilde{v}_{\mathrm{RA}}^*(s).\end{aligned}$$

The rest of the proof the follows the same as the previous case with T replaced by T'.

□

Corollary 5. *For all $s \in \mathcal{S}$, we have $V_{\mathrm{RAA}}^*(s) = v_{\mathrm{RAA}}^*(s)$.*

Proof of Theorem 1. Theorem 1 is now a direct consequence of the previous corollary together with Corollary 4 and Lemma 4. □

B PROOF OF RR MAIN THEOREM

We first define the value functions, $V^*_{\text{R1}}, V^*_{\text{R2}}, \tilde{V}^*_{\text{R}}, V^*_{\text{RR}} : \mathcal{S} \to \mathbb{R}$ by

$$\begin{aligned}
V^*_{\text{R1}}(s) &= \max_{\pi \in \Pi} \max_{\tau \in \mathbb{N}} r_1\left(\xi^\pi_s(\tau)\right), \\
V^*_{\text{R2}}(s) &= \max_{\pi \in \Pi} \max_{\tau \in \mathbb{N}} r_2\left(\xi^\pi_s(\tau)\right), \\
\tilde{V}^*_{\text{R}}(s) &= \max_{\pi \in \Pi} \max_{\tau \in \mathbb{N}} r_{\text{RR}}\left(\xi^{\mathbf{a}}_s(\tau)\right), \\
V^*_{\text{RR}}(s) &= \max_{\pi \in \Pi} \min\left\{\max_{\tau \in \mathbb{N}} r_1\left(\xi^\pi_s(\tau)\right), \max_{\tau \in \mathbb{N}} r_2\left(\xi^\pi_s(\tau)\right)\right\}.
\end{aligned}$$

We next define the value functions, $v^*_{\text{R1}}, v^*_{\text{R2}}, \tilde{v}^*_{\text{R}}, v^*_{\text{RR}} : \mathcal{S} \to \mathbb{R}$, which maximize over action sequences rather than policies:

$$\begin{aligned}
v^*_{\text{R1}}(s) &= \max_{\mathbf{a} \in \mathbb{A}} \max_{\tau \in \mathbb{N}} r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), \\
v^*_{\text{R2}}(s) &= \max_{\mathbf{a} \in \mathbb{A}} \max_{\tau \in \mathbb{N}} r_2\left(\xi^{\mathbf{a}}_s(\tau)\right), \\
\tilde{v}^*_{\text{R}}(s) &= \max_{\mathbf{a} \in \mathbb{A}} \max_{\tau \in \mathbb{N}} r_{\text{RR}}\left(\xi^{\mathbf{a}}_s(\tau)\right), \\
v^*_{\text{RR}}(s) &= \max_{\mathbf{a} \in \mathbb{A}} \min\left\{\max_{\tau \in \mathbb{N}} r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\tau \in \mathbb{N}} r_2\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\},
\end{aligned}$$

where r_{RR} is as in Theorem 2. Observe that for each $s \in \mathcal{S}$,

$$v^*_{\text{R1}}(s) \geq V^*_{\text{R1}}(s), \quad v^*_{\text{R2}}(s) \geq V^*_{\text{R2}}(s), \quad \tilde{v}^*_{\text{R}}(s) \geq \tilde{V}^*_{\text{R}}(s), \quad v^*_{\text{RR}}(s) \geq V^*_{\text{RR}}(s).$$

We now prove a series of lemmas that will be useful in the proof of the main theorem.

Lemma 6. *There are $\pi_1, \pi_2 \in \Pi$ such that*

$$v^*_{\text{R1}}(s) = \max_{\tau \in \mathbb{N}} r_1\left(\xi^{\pi_1}_s(\tau)\right) \text{ and } v^*_{\text{R2}}(s) = \max_{\tau \in \mathbb{N}} r_2\left(\xi^{\pi_2}_s(\tau)\right)$$

for all $s \in \mathcal{S}$.

Proof. We will just prove the result for $v^*_{\text{R1}}(s)$ since the other result follows identically. For each $s \in \mathcal{S}$, let τ_s be the smallest element of $\mathbb{N}$ for which

$$\max_{\mathbf{a} \in \mathbb{A}} r_1\left(\xi^{\mathbf{a}}_s(\tau_s)\right) = v^*_{\text{R1}}(s).$$

Moreover, for each $s \in \mathcal{S}$, let $\mathbf{a}_s$ be such that

$$r_1\left(\xi^{\mathbf{a}_s}_s(\tau_s)\right) = v^*_{\text{R1}}(s).$$

Let $\pi_1 \in \Pi$ be given by $\pi_1(s) = \mathbf{a}_s(0)$. It suffices to show that

$$r_1\left(\xi^{\pi_1}_s(\tau_s)\right) = v^*_{\text{R1}}(s) \tag{29}$$

for all $s \in \mathcal{S}$, for in this case, we have

$$v^*_{\text{R1}}(s) \geq \max_{\tau \in \mathbb{N}} r_1\left(\xi^{\pi_1}_s(\tau)\right) \geq r_1\left(\xi^{\pi_1}_s(\tau_s)\right) = v^*_{\text{R1}}(s) \quad \forall s \in \mathcal{S}.$$

We show (29) holds for each $s \in \mathcal{S}$ by induction on τ_s. First, suppose that $s \in \mathcal{S}$ is such that $\tau_s = 0$. Then

$$r_1\left(\xi^{\pi_1}_s(\tau_s)\right) = r_1(s) = r_1\left(\xi^{\mathbf{a}_s}_s(\tau_s)\right) = v^*_{\text{R1}}(s).$$

For the induction step, let $n \in \mathbb{N}$ and suppose that

$$r_1\left(\xi^{\pi_1}_s(\tau_s)\right) = v^*_{\text{R1}}(s) \quad \forall s \in \mathcal{S} \text{ such that } \tau_s \leq n.$$

Now fix some $x \in \mathcal{S}$ such that $\tau_x = n+1$. Notice that

$$\begin{aligned} v^*_{\mathrm{R}1}(x) &\geq v^*_{\mathrm{R}1}\left(f\left(x, \pi_1(x)\right)\right) \\ &\geq \max_{\mathbf{a}\in\mathbb{A}} r_1\left(\xi^{\mathbf{a}}_{f(x,\pi_1(x))}(n)\right) \\ &\geq r_1\left(\xi^{\mathbf{a}_x|_1}_{f(x,\pi_1(x))}(n)\right) \\ &= r_1\left(\xi^{[\pi_1(x),\mathbf{a}_x|_1]}_{x}(n+1)\right) \\ &= r_1\left(\xi^{\mathbf{a}_x}_{x}(\tau_x)\right) \\ &= v^*_{\mathrm{R}1}(x), \end{aligned}$$

so that $v^*_{\mathrm{R}1}\left(f\left(x, \pi_1(x)\right)\right) = v^*_{\mathrm{R}1}(x)$ and $\tau_{f(x,\pi_1(x))} \leq n$. It suffices to show

$$\tau_{f(x,\pi_1(x))} = n, \tag{30}$$

for then, by the induction assumption, we have

$$r_1\left(\xi^{\pi_1}_x(\tau_x)\right) = r_1\left(\xi^{\pi_1}_{f(x,\pi_1(x))}(n)\right) = v^*_{\mathrm{R}1}\left(f\left(x, \pi_1(x)\right)\right) = v^*_{\mathrm{R}1}(x).$$

To show (30), assume instead that

$$\tau_{f(x,\pi_1(x))} < n.$$

But

$$\begin{aligned} v^*_{\mathrm{R}1}(x) &\geq \max_{\mathbf{a}\in\mathbb{A}} r_1\left(\xi^{\mathbf{a}}_x\left(\tau_{f(x,\pi_1(x))}+1\right)\right) \\ &\geq r_1\left(\xi^{\left[\pi_1(x),\mathbf{a}_{f(x,\pi_1(x))}\right]}_x\left(\tau_{f(x,\pi_1(x))}+1\right)\right) \\ &= r_1\left(\xi^{\mathbf{a}_{f(x,\pi_1(x))}}_{f(x,\pi_1(x))}\left(\tau_{f(x,\pi_1(x))}\right)\right) \\ &= v^*_{\mathrm{R}1}\left(f\left(x, \pi_1(x)\right)\right) \\ &= v^*_{\mathrm{R}1}(x), \end{aligned}$$

so that

$$v^*_{\mathrm{R}1}(x) = \max_{\mathbf{a}\in\mathbb{A}} r_1\left(\xi^{\mathbf{a}}_x\left(\tau_{f(x,\pi_1(x))}+1\right)\right)$$

and thus

$$\tau_x \leq \tau_{f(x,\pi_1(x))} + 1 < n+1,$$

giving our desired contradiction. □

Corollary 6. *For all $s \in \mathcal{S}$, we have $V^*_{\mathrm{R}1}(s) = v^*_{\mathrm{R}1}(s)$ and $V^*_{\mathrm{R}2}(s) = v^*_{\mathrm{R}2}(s)$.*

Lemma 7. *There is a $\pi \in \Pi$ such that*

$$\tilde{v}^*_{\mathrm{R}}(s) = \max_{\tau\in\mathbb{N}} r_{\mathrm{RR}}\left(\xi^{\pi}_s(\tau)\right).$$

for all $s \in \mathcal{S}$.

Proof. This lemma follows by precisely the same proof as the previous lemma, with r_1, $v^*_{\mathrm{R}1}$, and π_1 replaced with r_{RR}, $\tilde{v}^*_{\mathrm{R}}$, and π respectively. □

Corollary 7. *For all $s \in \mathcal{S}$, we have $\tilde{V}^*_{\mathrm{R}}(s) = \tilde{v}^*_{\mathrm{R}}(s)$.*

Lemma 8. *Let $\zeta_1 : \mathbb{N} \to \mathbb{R}$ and $\zeta_2 : \mathbb{N} \to \mathbb{R}$. Then*

$$\begin{aligned} &\sup_{\tau\in\mathbb{N}} \max\left\{\min\left\{\zeta_1(\tau), \sup_{\tau'\in\mathbb{N}} \zeta_2(\tau+\tau')\right\}, \min\left\{\sup_{\tau'\in\mathbb{N}} \zeta_1(\tau+\tau'), \zeta_2(\tau)\right\}\right\} \\ &\quad = \min\left\{\sup_{\tau\in\mathbb{N}} \zeta_1(\tau), \sup_{\tau\in\mathbb{N}} \zeta_2(\tau)\right\}. \end{aligned}$$

Proof. We proceed by showing both inequalities corresponding to the above equality hold.

($\leq$) Observe that

$$\begin{aligned}
&\sup_{\tau\in\mathbb{N}} \max\left\{\min\left\{\zeta_1(\tau), \sup_{\tau'\in\mathbb{N}} \zeta_2(\tau+\tau')\right\}, \min\left\{\sup_{\tau'\in\mathbb{N}} \zeta_1(\tau+\tau'), \zeta_2(\tau)\right\}\right\} \\
&\quad\leq \max\left\{\min\left\{\sup_{\tau\in\mathbb{N}} \zeta_1(\tau), \sup_{\tau\in\mathbb{N}}\sup_{\tau'\in\mathbb{N}} \zeta_2(\tau+\tau')\right\}, \min\left\{\sup_{\tau\in\mathbb{N}}\sup_{\tau'\in\mathbb{N}} \zeta_1(\tau+\tau'), \sup_{\tau\in\mathbb{N}} \zeta_2(\tau)\right\}\right\} \\
&\quad= \min\left\{\sup_{\tau\in\mathbb{N}} \zeta_1(\tau), \sup_{\tau\in\mathbb{N}} \zeta_2(\tau)\right\}
\end{aligned}$$

($\geq$) Fix $\varepsilon > 0$. Choose $\tau_1, \tau_2 \in \mathbb{N}$ such that $\zeta_1(\tau_1) \geq \sup_{\tau\in\mathbb{N}} \zeta_1(\tau) - \varepsilon$ and $\zeta_2(\tau_2) \geq \sup_{\tau\in\mathbb{N}} \zeta_2(\tau) - \varepsilon$. Without loss of generality, we can assume $\tau_1 \leq \tau_2$. Then

$$\begin{aligned}
&\sup_{\tau\in\mathbb{N}} \max\left\{\min\left\{\zeta_1(\tau), \sup_{\tau'\in\mathbb{N}} \zeta_2(\tau+\tau')\right\}, \min\left\{\sup_{\tau'\in\mathbb{N}} \zeta_1(\tau+\tau'), \zeta_2(\tau)\right\}\right\} \\
&\quad\geq \sup_{\tau\in\mathbb{N}} \min\left\{\zeta_1(\tau), \sup_{\tau'\in\mathbb{N}} \zeta_2(\tau+\tau')\right\} \\
&\quad\geq \min\left\{\zeta_1(\tau_1), \sup_{\tau'\in\mathbb{N}} \zeta_2(\tau_1+\tau')\right\} \\
&\quad\geq \min\left\{\zeta_1(\tau_1), \zeta_2(\tau_2)\right\} \\
&\quad\geq \min\left\{\sup_{\tau\in\mathbb{N}} \zeta_1(\tau) - \varepsilon, \sup_{\tau\in\mathbb{N}} \zeta_2(\tau) - \varepsilon\right\} \\
&\quad= \min\left\{\sup_{\tau\in\mathbb{N}} \zeta_1(\tau), \sup_{\tau\in\mathbb{N}} \zeta_2(\tau)\right\} - \varepsilon.
\end{aligned}$$

But since $\varepsilon > 0$ was arbitrary, the desired inequality follows.

□

Lemma 9. *For each $s \in \mathcal{S}$,*

$$\tilde{v}^*_{\mathrm{R}}(s) = v^*_{\mathrm{RR}}(s).$$

Proof. For each $s \in \mathcal{S}$,

$$\begin{aligned}
\tilde{v}^*_{\mathrm{R}}(s) &= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} r_{\mathrm{RR}}\left(\xi^{\mathbf{a}}_s(\tau)\right) && (31)\\
&= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max\left\{\min\left\{r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), v^*_{\mathrm{R2}}\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\}, \min\left\{v^*_{\mathrm{R1}}\left(\xi^{\mathbf{a}}_s(\tau)\right), r_2\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\}\right\} && (32)\\
&= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max\left\{\min\left\{r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\mathbf{a}'\in\mathbb{A}} \max_{\tau'\in\mathbb{N}} r_2\left(\xi^{\mathbf{a}'}_{\xi^{\mathbf{a}}_s(\tau)}(\tau')\right)\right\},\right.\\
&\qquad\left.\min\left\{\max_{\mathbf{a}'\in\mathbb{A}} \max_{\tau'\in\mathbb{N}} r_1\left(\xi^{\mathbf{a}'}_{\xi^{\mathbf{a}}_s(\tau)}(\tau')\right), r_2\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\}\right\}\\
&= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max\left\{\min\left\{r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\mathbf{a}'\in\mathbb{A}} \max_{\tau'\in\mathbb{N}} r_2\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\tau')\right)\right\},\right.\\
&\qquad\left.\min\left\{\max_{\mathbf{a}'\in\mathbb{A}} \max_{\tau'\in\mathbb{N}} r_1\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\tau')\right), r_2\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\}\right\}\\
&= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max_{\mathbf{a}'\in\mathbb{A}} \max\left\{\min\left\{r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\tau'\in\mathbb{N}} r_2\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\tau')\right)\right\},\right.\\
&\qquad\left.\min\left\{\max_{\tau'\in\mathbb{N}} r_1\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\tau')\right), r_2\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\}\right\}\\
&= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max_{\mathbf{a}'\in\mathbb{A}} \max\left\{\min\left\{r_1\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau)\right), \max_{\tau'\in\mathbb{N}} r_2\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\tau')\right)\right\},\right.\\
&\qquad\left.\min\left\{\max_{\tau'\in\mathbb{N}} r_1\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau+\tau')\right), r_2\left(\xi^{[\mathbf{a},\mathbf{a}']_\tau}_s(\tau)\right)\right\}\right\} && (33)\\
&= \max_{\mathbf{a}\in\mathbb{A}} \max_{\tau\in\mathbb{N}} \max\left\{\min\left\{r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\tau'\in\mathbb{N}} r_2\left(\xi^{\mathbf{a}}_s(\tau+\tau')\right)\right\},\right.\\
&\qquad\left.\min\left\{\max_{\tau'\in\mathbb{N}} r_1\left(\xi^{\mathbf{a}}_s(\tau+\tau')\right), r_2\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\}\right\} && (34)\\
&= \max_{\mathbf{a}\in\mathbb{A}} \min\left\{\max_{\tau\in\mathbb{N}} r_1\left(\xi^{\mathbf{a}}_s(\tau)\right), \max_{\tau\in\mathbb{N}} r_2\left(\xi^{\mathbf{a}}_s(\tau)\right)\right\} && (35)\\
&= v^*_{\mathrm{RR}}(s),
\end{aligned}$$

where the equality between 31 and 32 follows from Corollary 6, the equality between 33 and 34 follows from Lemma 3, and the equality between 34 and 35 follows from Lemma 8. □

Before the next lemma, we need to introduce two last pieces of notation. First, we let $\overline{\Pi}$ be the set of augmented policies $\bar{\pi} : \mathcal{S} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{A}$, as in the previous section, but where

$$\mathcal{Y} = \{r_1(s) \mid s \in \mathcal{S}\} \quad \text{and} \quad \mathcal{Z} = \{r_2(s) \mid s \in \mathcal{S}\}.$$

Next, given $s \in \mathcal{S}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, and $\bar{\pi} \in \overline{\Pi}$, we let $\bar{\xi}^{\bar{\pi}}_s : \mathbb{N} \to \mathcal{S}$, $\bar{\eta}^{\bar{\pi}}_s : \mathbb{N} \to \mathcal{Y}$, and $\bar{\zeta}^{\bar{\pi}}_s : \mathbb{N} \to \mathcal{Z}$, be the solution of the evolution

$$\begin{aligned}
\bar{\xi}^{\bar{\pi}}_s(t+1) &= f\left(\bar{\xi}^{\bar{\pi}}_s(t), \bar{\pi}\left(\bar{\xi}^{\bar{\pi}}_s(t), \bar{\eta}^{\bar{\pi}}_s(t), \bar{\zeta}^{\bar{\pi}}_s(t)\right)\right),\\
\bar{\eta}^{\bar{\pi}}_s(t+1) &= \max\left\{r_1\left(\bar{\xi}^{\bar{\pi}}_s(t+1)\right), \bar{\eta}^{\bar{\pi}}_s(t)\right\},\\
\bar{\zeta}^{\bar{\pi}}_s(t+1) &= \max\left\{r_2\left(\bar{\xi}^{\bar{\pi}}_s(t+1)\right), \bar{\zeta}^{\bar{\pi}}_s(t)\right\},
\end{aligned}$$

for which $\bar{\xi}^{\bar{\pi}}_s(0) = s$, $\bar{\eta}^{\bar{\pi}}_s(0) = r_1(s)$, and $\bar{\xi}^{\bar{\pi}}_s(0) = r_2(s)$.

Lemma 10. *There is a $\bar{\pi} \in \overline{\Pi}$ such that*

$$v^*_{\mathrm{RR}}(s) = \min\left\{\max_{\tau\in\mathbb{N}} r_1\left(\bar{\xi}^{\bar{\pi}}_s(\tau)\right), \max_{\tau\in\mathbb{N}} r_2\left(\bar{\xi}^{\bar{\pi}}_s(\tau)\right)\right\}$$

for all $s \in \mathcal{S}$.

Proof. By Lemmas 6 and 7 together with Corollary 6, we can choose $\pi, \theta_1, \theta_2 \in \Pi$ such that

$$v^*_{\mathrm{R}1}(s) = \max_{\tau\in\mathbb{N}} r_1\left(\xi_s^{\theta_1}(\tau)\right) \quad \forall s \in \mathcal{S},$$
$$v^*_{\mathrm{R}2}(s) = \max_{\tau\in\mathbb{N}} r_2\left(\xi_s^{\theta_2}(\tau)\right) \quad \forall s \in \mathcal{S},$$
$$\tilde{v}^*_{\mathrm{R}}(s) = \max_{\tau\in\mathbb{N}} \max\left\{\min\left\{r_1\left(\xi_s^{\pi}(\tau)\right), v^*_{\mathrm{R}2}\left(\xi_s^{\pi}(\tau)\right)\right\}, \min\left\{r_2\left(\xi_s^{\pi}(\tau)\right), v^*_{\mathrm{R}1}\left(\xi_s^{\pi}(\tau)\right)\right\}\right\} \quad \forall s \in \mathcal{S}.$$

Define $\bar{\pi} \in \overline{\Pi}$ by

$$\bar{\pi}(s,y,z) = \begin{cases} \pi(s) & \max\{y,z\} < \tilde{v}^*_{\mathrm{R}}(s) \\ \theta_1(s) & \max\{y,z\} \geq \tilde{v}^*_{\mathrm{R}}(s) \text{ and } y \leq z, \\ \theta_2(s) & \max\{y,z\} \geq \tilde{v}^*_{\mathrm{R}}(s) \text{ and } y > z. \end{cases}$$

Now fix some $s \in \mathcal{S}$. For all $t \in \mathbb{N}$, set $\bar{x}_t = \bar{\xi}^{\bar{\pi}}_s(t)$, $\bar{y}_t = \bar{\eta}^{\bar{\pi}}_s(t) = \max_{\tau\leq t} r_1(\bar{x}_\tau)$, and $\bar{z}_t = \bar{\zeta}^{\bar{\pi}}_s(t) = \max_{\tau\leq t} r_2(\bar{x}_\tau)$, and also set $x^\circ_t = \xi^\pi_s(t)$. It suffices to show

$$v^*_{\mathrm{RR}}(s) \leq \min\left\{\max_{\tau\in\mathbb{N}} r_1\left(\bar{x}_\tau\right), \max_{\tau\in\mathbb{N}} r_2\left(\bar{x}_\tau\right)\right\}, \tag{36}$$

since the reverse inequality is immediate. We proceed in three steps.

1. We claim there exists a $t \in \mathbb{N}$ such that $\max\{r_1(\bar{x}_t), r_2(\bar{x}_t)\} \geq \tilde{v}^*_{\mathrm{R}}(\bar{x}_t)$.

   Suppose otherwise. Then $\bar{\pi}(\bar{x}_t, \bar{y}_t, \bar{z}_t) = \pi(\bar{x}_t)$ so that $\bar{x}_t = x^\circ_t$ for all $t \in \mathbb{N}$. Thus

   $$\begin{aligned}\max_{t\in\mathbb{N}} \max\{r_1(\bar{x}_t), r_2(\bar{x}_t)\} &< \max_{t\in\mathbb{N}} \tilde{v}^*_{\mathrm{R}}(\bar{x}_t) \\ &= \tilde{v}^*_{\mathrm{R}}(s) \\ &= \max_{\tau\in\mathbb{N}} \max\left\{\min\left\{r_1\left(x^\circ_\tau\right), v^*_{\mathrm{R}2}\left(x^\circ_\tau\right)\right\}, \min\left\{r_2\left(x^\circ_\tau\right), v^*_{\mathrm{R}1}\left(x^\circ_\tau\right)\right\}\right\} \\ &= \max_{\tau\in\mathbb{N}} \max\left\{\min\left\{r_1\left(\bar{x}_\tau\right), v^*_{\mathrm{R}2}\left(\bar{x}_\tau\right)\right\}, \min\left\{r_2\left(\bar{x}_\tau\right), v^*_{\mathrm{R}1}\left(\bar{x}_\tau\right)\right\}\right\} \\ &\leq \max_{\tau\in\mathbb{N}} \max\left\{r_1(\bar{x}_\tau), r_2(\bar{x}_\tau)\right\},\end{aligned}$$

   providing the desired contradiction.

2. Let T be the smallest element of $\mathbb{N}$ for which

   $$\max\left\{r_1(\bar{x}_T), r_2(\bar{x}_T)\right\} \geq v^*_{\mathrm{R}}(\bar{x}_T),$$

   which must exist by the previous step, and let T' be the smallest element of $\mathbb{N}$ for which

   $$\max\left\{\min\left\{r_1\left(x^\circ_{T'}\right), v^*_{\mathrm{R}2}\left(x^\circ_{T'}\right)\right\}, \min\left\{r_2\left(x^\circ_{T'}\right), v^*_{\mathrm{R}1}\left(x^\circ_{T'}\right)\right\}\right\} = \tilde{v}^*_{\mathrm{R}}(s),$$

   which must exist by our choice of π. We claim $T' \geq T$.

   Suppose otherwise. Since $\bar{x}_t = x^\circ_t$ for all $t \leq T$, then in particular $\bar{x}_{T'} = x^\circ_{T'}$, so that

   $$\max\left\{\min\left\{r_1\left(\bar{x}_{T'}\right), v^*_{\mathrm{R}2}\left(\bar{x}_{T'}\right)\right\}, \min\left\{r_2\left(\bar{x}_{T'}\right), v^*_{\mathrm{R}1}\left(\bar{x}_{T'}\right)\right\}\right\} = \tilde{v}^*_{\mathrm{R}}(s).$$

   But then

   $$\max\{r_1(\bar{x}_{T'}), r_2(\bar{x}_{T'})\} \geq \tilde{v}^*_{\mathrm{R}}(s) \geq \tilde{v}^*_{\mathrm{R}}(\bar{x}_{T'}).$$

   By our choice of T, we then have $T \leq T'$, creating a contradiction.

3. It follows from the previous step that

   $$\tilde{v}^*_{\mathrm{R}}(\bar{x}_T) = \tilde{v}^*_{\mathrm{R}}(x^\circ_T) = \tilde{v}^*_{\mathrm{R}}(s).$$

   By our choice of T, there are two cases: $r_1(\bar{x}_T) \geq \tilde{v}^*_{\mathrm{R}}(\bar{x}_T)$ and $r_2(\bar{x}_T) \geq \tilde{v}^*_{\mathrm{R}}(\bar{x}_T)$. We assume the first case and prove the desired result, with case two following identically. To reach a contradiction, assume

   $$r_2(\bar{x}_t) < \tilde{v}^*_{\mathrm{R}}(\bar{x}_T) \quad \forall t \in \mathbb{N}.$$

But then $\bar{\pi}(\bar{x}_t, \bar{y}_t, \bar{z}_t) = \theta_2(\bar{x}_t)$ for all $t \geq T$, so $v^*_{\text{R2}}(\bar{x}_T) = \max_{t \geq T} r_2(\bar{x}_t) < \tilde{v}^*_{\text{R}}(\bar{x}_T) \leq \tilde{v}^*_{\text{R}}(s)$. Thus $r_2(x^\circ_{T'}) \leq v^*_{\text{R2}}(x^\circ_{T'}) \leq v^*_{\text{R2}}(x^\circ_T) = v^*_{\text{R2}}(\bar{x}_T) < \tilde{v}^*_{\text{R}}(s)$. It follows that

$$\max\left\{\min\left\{r_1\left(x^\circ_{T'}\right), v^*_{\text{R2}}\left(x^\circ_{T'}\right)\right\}, \min\left\{r_2\left(x^\circ_{T'}\right), v^*_{\text{R1}}\left(x^\circ_{T'}\right)\right\}\right\} < \tilde{v}^*_{\text{R}}(s),$$

contradicting our choice of T'.

Thus $r_2(\bar{x}_t) \geq \tilde{v}^*_{\text{R}}(\bar{x}_T) = \tilde{v}^*_{\text{R}}(s)$ for some $t \in \mathbb{N}$ and also $r_1(\bar{x}_T) \geq \tilde{v}^*_{\text{R}}(\bar{x}_T) = \tilde{v}^*_{\text{R}}(s)$, so that (36) must hold by Lemma 9.

□

Corollary 8. *For all $s \in \mathcal{S}$, we have $V^*_{\text{RR}}(s, r_1(s), r_2(s)) = v^*_{\text{RR}}(s)$.*

Proof of Theorem 2. The proof of this theorem immediately follows from the previous corollary together with Corollary 7 and Lemma 9. □

C Proof of Optimality Theorem

Proof of Theorem 3. The inequalities in both lines of the theorem follow from the fact that for each $\pi \in \Pi$, we can define a corresponding augmented policy $\bar{\pi} \in \overline{\Pi}$ by

$$\bar{\pi}(s, y, z) = \pi(s) \quad \forall s \in \mathcal{S}, y \in \mathcal{Y}, z \in \mathcal{Z},$$

in which case $V^\pi_{\text{RAA}}(s) = V^{\bar{\pi}}_{\text{RAA}}(s)$ and $V^\pi_{\text{RR}}(s) = V^{\bar{\pi}}_{\text{RR}}(s)$ for each $s \in \mathcal{S}$. Note that in general, we cannot define a corresponding policy for each augmented policy, so the reverse inequality does not generally hold (see Figure 3 for intuition regarding this fact).

The equalities in both lines of the theorem are simply restatements of Lemma 5 and Lemma 9. □

D The SRABE and its Policy Gradient

Proof of Proposition 1. We here closely follow the proof of Theorem 3 in (4), which itself modifies the proofs of the Policy Gradient Theorems in Chapter 13.2 and 13.6 (60). We only make the minimal modifications required to adapt the PPO algorithm developed previously for the SRBE to on for the SRABE.

$$\begin{aligned}
\nabla_\theta \tilde{V}^{\pi_\theta}_{\text{RAA}}(s) =& \nabla_\theta \left(\sum_{a \in \mathcal{A}} \pi_\theta(a|s) \tilde{Q}^{\pi_\theta}_{\text{RAA}}(s,a) \right) \\
=& \sum_{a \in \mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) \tilde{Q}^{\pi_\theta}_{\text{RAA}}(s,a) \\
& \qquad + \pi_\theta(a|s) \nabla_\theta \min\left\{ \max\left\{ \tilde{V}^\pi_{\text{RAA}}\left(f(s,a)\right), r_{\text{RAA}}(s) \right\}, q(s) \right\} \Big) \\
=& \sum_{a \in \mathcal{A}} \Big(\nabla_\theta \pi_\theta(a|s) \tilde{Q}^{\pi_\theta}_{\text{RAA}}(s,a) \\
& \qquad + \pi_\theta(a|s) \left[q(s) < \tilde{V}_{\text{RAA}}\left(f(s,a)\right) < r_{\text{RAA}}(s) \right] \nabla_\theta \tilde{V}^\pi_{\text{RAA}}\left(f(s,a)\right) \Big) \qquad (37) \\
=& \sum_{s' \in \mathcal{S}} \left[\left(\sum_{k=0}^\infty \Pr(s \to s', k, \pi) \right) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s') \tilde{Q}^{\pi_\theta}_{\text{RAA}}(s',a) \right] \qquad (38) \\
=& \sum_{s' \in \mathcal{S}} \left[\left(\sum_{k=0}^\infty \Pr(s \to s', k, \pi) \right) \sum_{a \in \mathcal{A}} \pi_\theta(a|s') \frac{\nabla_\theta \pi_\theta(a|s')}{\pi_\theta(a|s')} \tilde{Q}^{\pi_\theta}_{\text{RAA}}(s',a) \right] \\
=& \sum_{s' \in \mathcal{S}} \left[\left(\sum_{k=0}^\infty \Pr(s \to s', k, \pi) \right) \mathbb{E}_{a \sim \pi_\theta(s')} \left[\nabla_\theta \ln \pi_\theta(a|s') \tilde{Q}^{\pi_\theta}_{\text{RAA}}(s',a) \right] \right] \\
\propto& \mathbb{E}_{s' \sim d'_\pi(s)} \mathbb{E}_{a \sim \pi_\theta(s')} \left[\nabla_\theta \ln \pi_\theta(a|s') \tilde{Q}^{\pi_\theta}_{\text{RAA}}(s',a) \right],
\end{aligned}$$

where the equality between (37) and (38) comes from rolling out the term $\nabla_\theta \tilde{V}^{\pi}_{\text{RAA}}(f(s,a))$ (see Chapter 13.2 in (60) for details), and where $\Pr(s \to s', k, \pi)$ is the probability that under the policy π, the system is in state s' at time k given that it is in state s at time 0. □

Note, Proposition 1 is vital to updating the actor in Algorithm 1.

E THE DO-HJ-PPO ALGORITHM

In this section, we outline the details of our Actor-Critic algorithm DO-HJ-PPO beyond the details given in Algorithm 1.

Algorithm 1 : DO-HJ-PPO (Actor-Critic)

Require: Composed and Decomposed Actor parameters θ and θ_i, Composed and Decomposed Critic parameters ω and ω_i, GAE λ, learning rate β_k and discount factor γ. Let B^γ amd B_i^γ represent the Bellman update and decomposed Bellman update for the users choice of problem (RR or RAA).

Define *Composed* Actor and Critic $\tilde{Q}$
Define *Decomposed* Actor(s) and Critic(s) $\tilde{Q}_i$
for $k = 0, 1, \ldots$ **do**
  for $t = 0$ to $T - 1$ **do**
    Sample trajectories for $\tau_t : \{\hat{s}_t, a_t, \hat{s}_{t+1}\}$
    Define $\tilde{\ell}(s_t)$ with Decomposed Critics $\tilde{Q}_i(s_t)$ (Theorems 1 & 2)
    Composed Critic update:

$$\omega \leftarrow \omega - \beta_k \nabla_\omega \tilde{Q}(\tau_t) \cdot \left(\tilde{Q}(\tau_t) - B^\gamma[\tilde{Q}, \tilde{r}](\tau_t)\right)$$

    Compute Bellman-GAE A^λ_{HJ} with B^γ
    (Standard) update Composed Actor
    Decomposed Critic update(s):

$$\omega \leftarrow \omega - \beta_k \nabla_\omega \tilde{Q}_i(\tau_t) \cdot \left(\tilde{Q}_i(\tau_t) - B_i^\gamma[\tilde{Q}_i](\tau_t)\right)$$

    Compute Bellman-GAE A^λ_i with B^γ_i
    (Standard) update Decomposed Actor(s)
  end for
end for
return parameter θ, ω

In Algorithm 1, the Bellman update $B^\gamma[\tilde{Q}, \tilde{r}]$ differs for the RAA task and RR task, and the $B_i^\gamma[\tilde{Q}]$ differs between the reach, avoid, and reach-avoid tasks. These Bellman updates are explicitly specified in the Supplementary Material.

E.1 THE SPECIAL BELLMAN UPDATES AND THE CORRESPONDING GAES

Akin to previous HJ-RL policy algorithms, namely RCPO (6), RESPO (3) and RCPPO (4), DO-HJ-PPO fundamentally depends on the discounted HJ Bellman updates (1). To solve the RAA and RR problems with the special rewards defined in Theorems 1 & 2, DO-HJ-PPO utilizes the Reach, Avoid and Reach-Avoid Bellman updates, given by

$$B^\gamma_R[Q \mid r](s,a) = (1-\gamma) r(s) + \gamma \max\{r(s), Q(s,a)\}, \tag{39}$$

$$B^\gamma_A[Q \mid q](s,a) = (1-\gamma) q(s) + \gamma \min\{q(s), Q(s,a)\}, \tag{40}$$

$$B^\gamma_{RA}[Q \mid r,q](s,a) = (1-\gamma)\min\{r(s), q(s)\} + \gamma \min\{q(s), \max\{r(s), Q(s,a)\}\}. \tag{41}$$

To improve our algorithm, we incorporate the Generalized Advantage Estimate corresponding to these Bellman equations in the updates of the Actors. As outlined in Section A of (4), the GAE may be defined with a reduction function corresponding to the appropriate Bellman function which

will be applied over a trajectory roll-out. We generalize the Reach GAE definition given in (4) to propose a Reach-Avoid GAE (the Avoid GAE is simply the flip of the Reach GAE) as all will be used in DO-HJ-PPO algorithm for either RAA or RR problems. Consider a reduction function $\phi_{RA}^{(n)} : \mathbb{R}^n \to \mathbb{R}$, defined by

$$\phi_{RA}^{(n)}(x_1, x_2, x_3, \ldots, x_{2n+1}) = \phi_{RA}^{(1)}(x_1, x_2, \phi_{RA}^{(n-1)}(x_3, \ldots, x_{2n+1})), \tag{42}$$

$$\phi_{RA}^{(1)}(x, y, z) = (1-\gamma)\min\{x, y\} + \gamma \min\{y, \max\{x, z\}\}. \tag{43}$$

The k-step Reach-Avoid Bellman advantage $A_{RA}^{\pi(k)}$ is then given by,

$$A_{RA}^{(k)}(s) = \phi_{RA}^{(n)}\left(r(s_t), q(s_t), \ldots, r(s_{t+k-1}), q(s_{t+k-1}), V^{(}s_{t+k})\right) - V^{(}s_{t+k}). \tag{44}$$

We may then define the Reach-Avoid GAE A_{RA}^{λ} as the λ-weighted sum over the advantage functions

$$A_{RA}^{\lambda}(s) = \frac{1}{1-\lambda}\sum_{k=1}^{\infty} \lambda^k A_{RA}^{(k)}(s) \tag{45}$$

which may be approximated over any finite trajectory sample. See (4) for further details.

E.2 Modifications from standard PPO

To address the RAA and RR problems, DO-HJ-PPO introduces several key modifications to the standard PPO framework (61):

Additional actor and critic networks are introduced to represent the decomposed objectives. Rather than learning the decomposed objectives separately from the composed objective, DO-HJ-PPO optimizes all objectives simultaneously. This design choice is motivated by two primary factors: (i) simplicity and minor computational speed-up, and (ii) coupling between the decomposed and composed objectives during learning.

The decomposed trajectories are initialized using states sampled from the composed trajectory, we refer to as *coupled resets*.
While it is possible to estimate the decomposed objectives independently—i.e., prior to solving the composed task—this approach might lead to inaccurate or irrelevant value estimates in on-policy settings. For example, in the RAA problem, the decomposed objective may prioritize avoiding penalties, while the composed task requires reaching a reward region without incurring penalties. In such a case, a decomposed policy trained in isolation might converge to an optimal strategy within a reward-irrelevant region, misaligned with the overall task. Empirically, we observe that omitting coupled resets causes DO-HJ-PPO to perform no better than standard baselines such as CPPO, whereas their inclusion significantly improves performance.

The special RAA and RR rewards are defined using the decomposed critic values and updated using their corresponding Bellman equations.
This procedure is directly derived from our theoretical results (Theorems 1 and 2), which establish the validity of using modified rewards within the respective RA and R Bellman frameworks. These rewards are used to compute the composed critic target as well as the actor's GAE. In Algorithm 1, this process is reflected in the critic and actor updates corresponding to the composed objective.

F DDQN Demonstration

As described in the paper, we demonstrate the novel RAA and RR problems in a 2D Q-learning problem where the value function may be observed easily. We juxtapose these solitons with those of the previously studied RA and R problems which consider more simple objectives. To solve all values, we employ the standard Double-Deep Q learning approach (DDQN) (66) with only the special Bellman updates.

F.1 GRID-WORLD ENVIRONMENT

The environment is taken from (2) and consists of two dimensions, $s = (x, y)$, and three actions, $a \in \{\text{left}, \text{straight}, \text{right}\}$, which allow the agent to maneuver through the space. The deterministic dynamics of the environment are defined by constant upward flow such that,

$$f((x_i, y_i), a_i) = \begin{cases} (x_{i-1}, y_{i+1}) & a_i = \text{left} \\ (x_i, y_{i+1}) & a_i = \text{straight} \\ (x_{i+1}, y_{i+1}) & a_i = \text{right} \end{cases} \tag{46}$$

and if the agent reaches the boundary of the space, defined by $x \geq |2|$, $y \leq -2$ and $y \geq 10$, the trajectory is terminated. The 2D space is divided into 80×120 cells which the agent traverses through.

In the RA and RAA experiments, the reward function r is defined as the negative signed-distance function to a box with dimensions $(x_c, y_c, w, h) = (0, 4.5, 2, 1.5)$, and thus is negative iff the agent is outside of the box. The penalty function q is defined as the minimum of three (positive) signed distance functions for boxes defined at $(x_c, y_c, w, h) = (\pm 0.75, 3, 1, 1)$ and $(x_c, y_c, w, h) = (0, 6, 2.5, 1)$, and thus is positive iff the agent is outside of all boxes.

In the R and RR experiments, one or two rewards are used. In the R experiment, the reward function r is defined as the maximum of two negative signed-distance function of boxes with dimensions $(x_c, y_c, w, h) = (\pm 1.25, 0, 0.5, 2)$, and thus is negative iff the agent is outside of both boxes. In the RR experiment, the rewards r_1 and r_2 are defined as the negative signed distance functions of the same two boxes independently, and thus are positive if the agent is in one box or the other respectively.

F.2 DDQN DETAILS

As per our theoretical results in Theorems 1 and 2, we may now perform DDQN to solve the RAA and RR problems with solely the previously studied Bellman updates for the RA (2) and R problems (1). We compare these solutions with those corresponding to the RA and R problems *without* the special RAA and RR targets, and hence solve the previously posed problems. For all experiments, we employ the same adapted algorithm as in (2), with no modification of the hyper-parameters given in Table 1.

Table 1: Hyperparameters for DDQN Grid World

DDQN hyperparameters	**Values**
Network Architecture	MLP
Numbers of Hidden Layers	2
Units per Hidden Layer	100, 20
Hidden Layer Activation Function	tanh
Optimizer	Adam
Discount factor γ	0.9999
Learning rate	1e-3
Replay Buffer Size	1e5 transitions
Replay Batch Size	100
Train-Collect Interval	10
Max Updates	4e6

G BASELINES

In both RAA and RR problems, we employ Constrained PPO (CPPO) (8) as the major baseline as it can handle secondary objectives which are reformulated as constraints. The algorithm was not designed to minimize its constraints necessarily but may do so in attempting to satisfy them. As a novel direction in RL, few algorithms have been designed to optimize max/min accumulated costs and thus CPPO serves as the best proxy. Below we also include a naively decomposed STL algorithm to offer some insight into direct approaches to optimizing the max/min accumulated reward.

G.1 CPPO BASELINES

Although CPPO formulations do not directly consider dual-objective optimization, the secondary objective in RAA (avoid penalty) or overall objective in RR (reach both rewards) may be transformed into constraints to be satisfied of a surrogate problem. For the RAA problem, this may be defined as

$$\max_\pi \mathbb{E}_\pi \left[\sum_t^\infty \gamma^t \max_{t' \le t} r(s^\pi_{t'}) \right] \quad \text{s.t.} \quad \min_t q(s^\pi_t) \ge 0. \tag{47}$$

For the RR problem, one might propose that the fairest comparison would be to formulate the surrogate problem in the same fashion, with achievement of both costs as a constraint, such that

$$\max_\pi \mathbb{E}_\pi \left[\sum_t^\infty \gamma^t \min \left\{ \max_{t' \le t} r_1(s^\pi_{t'}), \max_{t' \le t} r_2(s^\pi_{t'}) \right\} \right] \quad \text{s.t.} \quad \min \left\{ \max_t r_1(s^\pi_t), \max_t r_2(s^\pi_t) \right\} \ge 0, \tag{48}$$

which we define as variant 1 (CPPO-v1). Empirically, however, we found this formulation to be the poorest by far, perhaps due to the abundance of the non-smooth combinations. We thus also compare with more naive formulations which relax the outer minimizations to summation in the reward

$$\max_\pi \mathbb{E}_\pi \left[\sum_t^\infty \gamma^t \max_{t' \le t} r_1(s^\pi_{t'}) + \max_{t' \le t} r_2(s^\pi_{t'}) \right] \quad \text{s.t.} \quad \min \left\{ \max_t r_1(s^\pi_t), \max_t r_2(s^\pi_t) \right\} \ge 0, \tag{49}$$

which we define as variant 2 (CPPOv2), and additionally, in the constraint

$$\max_\pi \mathbb{E}_\pi \left[\sum_t^\infty \gamma^t \max_{t' \le t} r_1(s^\pi_{t'}) + \max_{t' \le t} r_2(s^\pi_{t'}) \right] \quad \text{s.t.} \quad \max_t r_1(s^\pi_t) + \max_t r_2(s^\pi_t) \ge 0, \tag{50}$$

which we define as variant 3 (CPPOv3). This last approach, although naive and seemingly unfair, vastly outperforms the other variants in the RR problem.

G.2 STL BASELINES

In contrast with constrained optimization, one might also incorporate the STL methods, which in the current context simply decompose and optimize the independent objectives. For the RAA problem, the standard RA solution serves as a trivial STL baseline since we may attempt to continuously attempt to reach the solution while avoiding the obstacle. In the RR case, we define a decomposed STL baseline (DSTL) which naively solves both R problems, and selects the one with lower value to achieve first.

H DETAILS OF RAA & RR EXPERIMENTS: HOPPER

The Hopper environment is taken from Gym (67) and (4). In both RAA and RR problems, we define rewards and penalties based on the position of the Hopper head, which we denote as (x, y) in this section.

In the RAA task, the reward is defined as

$$r(x, y) = \sqrt{\|x - 2\| + |y - 1.4\|} - 0.1 \tag{51}$$

to incentive the Hopper to reach its head to the position at $(x, y) = (2, 1.4)$. The penalty q is defined as the minimum of signed distance functions to a ceiling obstacle at $(1, 0)$, wall obstacles at $x > 2$ and $x < 0$ and a floor obstacle at $y < 0.5$. In order to safely arrive at high reward (and always avoid the obstacles), the Hopper thus must pass under the ceiling and not dive or fall over in the achievement of the target, as is the natural behavior.

In the RR task, the first reward is defined again as

$$r_1(x, y) = \sqrt{\|x - 2\| + |y - 1.4\|} - 0.1 \tag{52}$$

to incentive the Hopper to reach its head to the position at $(x, y) = (2, 1.4)$, and the second reward as

$$r_2(x, y) = \sqrt{\|x - 0\| + |y - 1.4\|} - 0.1 \tag{53}$$

to incentive the Hopper to reach its head to the position at $(x, y) = (0, 1.4)$. In order to achieve both rewards, the Hopper must thus hop both forwards and backwards without crashing or diving.

In all experiments, the Hopper is initialized in the default standing posture at a random $x \in [0, 2]$ so as to learn a position-agnostic policy. The DO-HJ-PPO parameters used to train these problems can be found in Table 2.

Table 2: Hyperparameters for Hopper Learning

Hyperparameters for DO-HJ-PPO	**Values**
Network Architecture	MLP
Units per Hidden Layer	256
Numbers of Hidden Layers	2
Hidden Layer Activation Function	tanh
Entropy coefficient	Linear Decay 1e-2 $\rightarrow$ 0
Optimizer	Adam
Discount factor γ	Linear Anneal 0.995 $\rightarrow$ 0.999
GAE lambda parameter	0.95
Clip Ratio	0.2
Actor Learning rate	Linear Decay 3e-4 $\rightarrow$ 0
Reward/Cost Critic Learning rate	Linear Decay 3e-4 $\rightarrow$ 0
Add'l Hyperparameters for CPPO	
K_P	1
K_I	1e-4
K_D	1

I Details of RAA & RR Experiments: F16

The F16 environment is taken from (4), including a F16 fighter jet with a 26 dimensional observation. The jet is limited to a flight corridor with up to 2000 relative position north (x_{PN}), 1200 relative altitude (x_H), and ± 500 relative position east (x_{PE}).

In the RAA task, the reward is defined as

$$r(x, y) = \frac{1}{5}|x_{PN} - 1500| - 50 \tag{54}$$

to incentivize the F16 to fly through the geofence defined by the vertical slice at 1500 relative position north. The penalty q is defined as the minimum of signed distance functions to geofence (wall) obstacles at $x_{PN} > 2000$ and $|x_{PE}| > 500$ and a floor obstacle at $x_H < 0$. In order to safely arrive at high reward (and always avoid the obstacles), the F16 thus must fly through the target geofence and then evade crashing into the wall directly in front of it.

In the RR task, the rewards are defined as

$$r_1(x_{PN}, x_H) = \frac{1}{5}\sqrt{\|x_{PN} - 1250\| + |y - 850\|} - 30 \tag{55}$$

and

$$r_2(x_{PN}, x_H) = \frac{1}{5}\sqrt{\|x_{PN} - 1250\| + |y - 350\|} - 30 \tag{56}$$

to incentive the F16 to reach both low and high-altitude horizontal cylinders. In order to achieve both rewards, the F16 must thus aggressively pitch, roll and yaw between the two targets.

In all experiments, the F16 is initialized with position $x_{PN} \in [250, 750], x_H \in [300, 900], x_{PE} \in [-250, 250]$ and velocity in $v \in [200, 450]$. Additionally, the roll, pitch, and yaw are initialized with $\pm\pi/16$ to simulate a variety of approaches to the flight corridor. Further details can be found in (4). The DO-HJ-PPO parameters used to train these problems can be found in Table 3.

Table 3: Hyperparameters for F16 Learning

Hyperparameters for DO-HJ-PPO	**Values**
Network Architecture	MLP
Units per Hidden Layer	256
Numbers of Hidden Layers	2
Hidden Layer Activation Function	tanh
Entropy coefficient	Linear Decay 1e-2 $\rightarrow$ 0
Optimizer	Adam
Discount factor γ	Linear Anneal 0.995 $\rightarrow$ 0.999
GAE lambda parameter	0.95
Clip Ratio	0.2
Actor Learning rate	Linear Decay 1e-3 $\rightarrow$ 0
Reward/Cost Critic Learning rate	Linear Decay 1e-3 $\rightarrow$ 0
Add'l Hyperparameters for CPPO	
K_P	1
K_I	1e-4
K_D	1

J BROADER IMPACTS

This paper touches on advancing fundamental methods for Reinforcement Learning. In particular, this work falls into the class of methods designed for Safe Reinforcement Learning. Methods in this class are primarily intended to prevent undesirable behaviors in virtual or cyber-physical systems, such as preventing crashes involving self-driving vehicles or potentially even unacceptable speech among chatbots. It is an unfortunate truth that safe learning methods can be repurposed for unintended use cases, such as to prevent a malicious agent from being captured, but the authors do not foresee the balance of potential beneficial and malicious applications of this method to be any greater than other typical methods in Safe Reinforcement Learning.

K ACKNOWLEDGMENTS

This section has been redacted for the purpose of anonymous review.