Bellman Optimality of Average-Reward Robust Markov Decision Processes with a Constant Gain

Shengbo Wang
Daniel J. Epstein Department of Industrial and Systems Engineering
University of Southern California
shengbow@usc.edu

Nian Si
Department of Industrial Engineering and Decision Analytics
Hong Kong University of Science and Technology
niansi@ust.hk

1 Introduction

In this paper, we consider optimality conditions for average-reward robust Markov decision processes (MDPs), where a controller aims to optimize the long-run average reward in the presence of an adversary. Specifically, we consider the value

$$\overline{\alpha}(\mu, \Pi, \mathrm{K}) := \sup_{\pi\in\Pi} \inf_{\kappa\in\mathrm{K}} \limsup_{n\to\infty} E_\mu^{\pi,\kappa} \frac{1}{n} \sum_{k=0}^{n-1} r(X_k, A_k),$$

where Π and K denote the controller's and adversary's policy classes, respectively, and μ is the initial distribution. We assume that the adversary is S-rectangular, and that both policy classes may be history-dependent (denoted by Π_{H} and K_{H}) or stationary (denoted by Π_{S} and K_{S}), with potentially asymmetric information structures between the controller and adversary decisions.

We are interested in identifying conditions under which $\overline{\alpha}(\mu, \Pi, \mathrm{K}) = \alpha^*$ holds for all initial distributions μ. Here, $\alpha^* \in [0, 1]$ arises as part of a solution pair (u^*, α^*) to the following robust Bellman equation with a constant gain:

$$u^*(s) = \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s, A_0) - \alpha^* + u^*(X_1)], \tag{1.1}$$

where $\mathcal{Q}$ and $\mathcal{P}_s$ are the controller's and adversary's decision sets, respectively.

Our main results can be summarized as follows.

- When the robust Bellman equation (1.1) admits a solution (u^*, α^*), α^* coincides with the optimal average-rewards $\alpha^* = \overline{\alpha}(\Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{H}}) = \overline{\alpha}(\Pi_{\mathrm{S}}, \mathrm{K}_{\mathrm{H}}) = \overline{\alpha}(\Pi_{\mathrm{S}}, \mathrm{K}_{\mathrm{S}})$, independent of the initial distribution (cf. Section 3.2). In particular, stationary controller policies are optimal for $\overline{\alpha}(\Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{H}})$. However, in general, $\overline{\alpha}(\Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{S}}) \neq \alpha^*$; (cf. Section 5).
- Theorem 4 in Section 4 implies that if the controller is communicating (as in Definition 3) and $\mathcal{Q}$ is compact, then (1.1) has a solution.
- Theorem 5 in Section 4 shows that if the adversary is communicating, $\mathcal{P}_s,\ s \in S$ are convex and compact, and $\mathcal{Q}$ is convex, then the Bellman equation (1.1) has a solution.
- Section 5 shows that if both the controller and the adversary are communicating and compact (not necessarily convex), and $\mathrm{K} = \mathrm{K}_{\mathrm{S}}$, then a stationary policy is optimal for the controller if and only if $\alpha' = \alpha^*$. Here, α' denotes the solution to (3.2), obtained by swapping the sup-inf order in (1.1).

1.1 Literature review

Robust MDPs: While the Bellman optimality of discounted-reward robust MDPs has been extensively studied [5, 9, 10, 7], the corresponding results for the average-reward setting remain

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

underexplored. To the best of our knowledge, Wang et al. [8] provides the first results under the stringent assumptions of SA-rectangularity and uniform unichains. Grand-Clement et al. [4] focus on Blackwell optimality, showing that ϵ-Blackwell optimal policies always exist under SA-rectangularity. Moreover, they demonstrate that in S-rectangular RMDPs, average-optimal policies may fail to exist; and even when they do exist, they may need to be strictly history-dependent.

Stochastic games (SGs): S-rectangular robust MDPs can be viewed as a generalization of zero-sum SGs. They extend the standard SGs framework in the following ways: (i) an asymmetry of information, where the controller may use history-dependent policies while the adversary is restricted to stationary or Markovian ones, and (ii) ambiguity-set constraints, where the adversary's feasible set may be nonconvex, in contrast to the convex mixed-strategy sets of SGs.

There is an extensive literature on average-payoff SGs; see Section 5 of Filar and Vrieze [3]. The present work, which establishes constant-gain Bellman optimality for robust MDPs under one-sided communication and compactness assumptions, also contributes to the SG literature.

2 Canonical Construction and the Optimal Robust Control Problem

In this section, we first present a brief but self-contained canonical construction of the probability space, the processes of interest, and the controller's and adversary's policy classes. The construction closely follows Wang et al. [7], to which we refer the reader for additional details.

Let S, A be finite state and action spaces, each equipped with the discrete Borel σ-fields $\mathcal{S}$ and $\mathcal{A}$, respectively. Define the underlying measurable space $(\Omega, \mathcal{F})$ with $\Omega = (S \times A)^{\mathbb{Z}_{\geq 0}}$ and $\mathcal{F}$ the corresponding cylinder σ-field. The process $\{(X_t, A_t), t \geq 0\}$ is defined by point evaluation, i.e., $X_t(\omega) = s_t$ and $A_t(\omega) = a_t$ for all $t \geq 0$ and any $\omega = (s_0, a_0, s_1, a_1, \ldots) \in \Omega$.

The history set $\mathbf{H}_t$ at time t contains all t-truncated sample paths $\mathbf{H}_t := \{h_t = (s_0, a_0, \ldots, a_{t-1}, s_t) : \omega = (s_0, a_0, s_1 \ldots) \in \Omega\}$. We also define the random element $H_t : \Omega \to \mathbf{H}_t$ by point evaluation $H_t(\omega) = h_t$, and the σ-field $\mathcal{H}_t := \sigma(H_t)$.

Given a prescribed subset $\mathcal{Q} \subset \mathcal{P}(\mathcal{A})$, a controller policy π is a sequence of decision rules $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ where each π_t is a measure-valued function $\pi_t : \mathbf{H}_t \to \mathcal{Q}$, represented in conditional distribution form as $\pi_t(a|h_t) \in [0, 1]$ with $\sum_{a \in A} \pi_t(a|h_t) = 1$. The history-dependent controller policy class is therefore $\Pi_{\mathrm{H}}(\mathcal{Q}) := \{\pi = (\pi_0, \pi_1, \ldots) : \pi_t \in \{\mathbf{H}_t \to \mathcal{Q}\}, \ \forall t \geq 0\}$.

A controller policy $\pi = (\pi_0, \pi_1, \ldots)$ is stationary if for any $t_1, t_2 \geq 0$ and $h_{t_1} \in \mathbf{H}_{t_1}, h'_{t_2} \in \mathbf{H}_{t_2}$ such that $s_{t_1} = s'_{t_2}$, we have $\pi_{t_1}(\cdot|h_{t_1}) = \pi_{t_2}(\cdot|h'_{t_2})$. In particular, this means $\pi_t(a|h_t) = \Delta(a|s_t)$ where $h_t = (s_0, a_0, \ldots, s_t)$ for some $\Delta : S \to \mathcal{Q}$ for all $t \geq 0$. Thus, a stationary controller policy can be identified with $\Delta : S \to \mathcal{Q}$, i.e., $\pi = (\Delta, \Delta, \ldots)$. Accordingly, the stationary policy class for the controller is $\Pi_{\mathrm{S}}(\mathcal{Q}) := \{(\Delta, \Delta, \ldots) : \Delta \in \{S \to \mathcal{Q}\}\}$, which is identified with $\{S \to \mathcal{Q}\}$.

On the adversary side, for each $s \in S$ we fix a prescribed set of measure-valued functions $\mathcal{P}_s \subset \{A \to \mathcal{P}(\mathcal{S})\}$. The product set $\mathcal{P} := \bigtimes_{s \in S} \mathcal{P}_s$ is called an S-rectangular ambiguity set.

Given $\{\mathcal{P}_s : s \in S\}$, a history-dependent S-rectangular adversary policy κ is a sequence of adversarial decision rules $\kappa = (\kappa_0, \kappa_1, \kappa_2, \ldots)$. Each decision rule κ_t specifies the conditional distribution of the next state given a history $h_t \in \mathbf{H}_t$ and an action $a \in A$, i.e., $\kappa_t(s'|h_t, a) \in [0, 1]$ with $\sum_{s' \in S} \kappa_t(s'|h_t, a) = 1$. The history-dependent adversary policy class is $\mathrm{K}_{\mathrm{H}}(\mathcal{P}) := \{\kappa = (\kappa_0, \kappa_1, \ldots) : \kappa_t(\cdot|h_t, \cdot) \in \mathcal{P}_{s_t}, \text{ where } h_t = (s_0, a_0, \ldots, s_t), \ \forall t \geq 0\}$.

Analogous to the controller side, a stationary adversary policy $\kappa = (\kappa_0, \kappa_1, \ldots)$ can be identified with $p \in \mathcal{P}$, i.e., $\kappa = (p, p, \ldots)$ with $\kappa_t(s'|h_t, a) = p(s'|s_t, a)$ where $h_t = (s_0, a_0, \ldots, s_t)$. Thus, the stationary adversary policy class is $\mathrm{K}_{\mathrm{S}}(\mathcal{P}) := \{(p, p, \ldots) : p \in \mathcal{P}\}$, which can be identified directly with $\mathcal{P}$.

As shown in Wang et al. [7], for $\Pi = \Pi_{\mathrm{H}}(\mathcal{Q})$ or $\Pi_{\mathrm{S}}(\mathcal{Q})$ and $\mathrm{K} = \mathrm{K}_{\mathrm{H}}(\mathcal{P})$ or $\mathrm{K}_{\mathrm{S}}(\mathcal{P})$ the triple $\mu \in \mathcal{P}(\mathcal{S}), \pi \in \Pi, \kappa \in \mathrm{K}$ uniquely defines a probability measure $P_\mu^{\pi,\kappa}$ on $(\Omega, \mathcal{F})$. The expectation under $P_\mu^{\pi,\kappa}$ is denoted by $E_\mu^{\pi,\kappa}$.

This paper considers the optimal robust control of the upper and lower long-run average rewards associated with a robust MDP instance $(\mathcal{Q}, \mathcal{P}, r)$ defined by

$$\overline{\alpha}(\mu, \Pi, \mathrm{K}) := \sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} \overline{\alpha}(\mu, \pi, \kappa) \ \text{ and } \ \underline{\alpha}(\mu, \Pi, \mathrm{K}) := \sup_{\pi \in \Pi} \inf_{\kappa \in \mathrm{K}} \underline{\alpha}(\mu, \pi, \kappa),$$

where

$$\overline{\alpha}(\mu,\pi,\kappa) := \limsup_{n\to\infty} E_\mu^{\pi,\kappa}\frac{1}{n}\sum_{k=0}^{n-1} r(X_k,A_k) \text{ and } \underline{\alpha}(\mu,\pi,\kappa) := \liminf_{n\to\infty} E_\mu^{\pi,\kappa}\frac{1}{n}\sum_{k=0}^{n-1} r(X_k,A_k).$$

The controller's policy class is either $\Pi_{\mathrm{H}}(\mathcal{Q})$ or $\Pi_{\mathrm{S}}(\mathcal{Q})$, while the adversary's policy class is either $\mathrm{K}_{\mathrm{H}}(\mathcal{P})$ or $\mathrm{K}_{\mathrm{S}}(\mathcal{P})$. For notational simplicity, we will suppress the dependence of Π and K on $\mathcal{Q}$ and $\mathcal{P}$ whenever it is clear from the context.

3 Robust Bellman Equations and Optimality

In this section, we define the constant-gain robust Bellman equation and show that its solution determines the long-run average reward of the robust control problem. This also implies stationary optimality for the controller in the $\overline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})$ and $\underline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})$ case.

3.1 Robust Bellman Equations with a Constant Gain

Definition 1. $(u^*,\alpha^*)\in\{S\to\mathbb{R}\}\times[0,1]$ is said to be a solution of the robust Bellman equation with a constant gain if

$$u^*(s) = \sup_{\phi\in\mathcal{Q}}\inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0)-\alpha^*+u^*(X_1)], \quad \forall s\in S. \tag{3.1}$$

Here, the expectation is taken w.r.t. the measure $P_{\phi,p_s}(A_0=a,X_1=s')=\phi(a)p_{s,a}(s')$. We say that $(u',\alpha')\in\{S\to\mathbb{R}\}\times[0,1]$ is a solution to the inf-sup equation with a constant gain if

$$u'(s) = \inf_{p_s\in\mathcal{P}_s}\sup_{\phi\in\mathcal{Q}} E_{\phi,p_s}[r(s,A_0)-\alpha'+u'(X_1)], \quad \forall s\in S. \tag{3.2}$$

It is useful to introduce the following discounted robust Bellman equation. These will serve as key theoretical tools in establishing the existence of solutions to the average-reward equation (3.1).

Definition 2. $v_\gamma^*: S\to\mathbb{R}$ is the unique solution of the γ-discounted robust Bellman equation if

$$v_\gamma^*(s) = \sup_{\phi\in\mathcal{Q}}\inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0)+\gamma v_\gamma^*(X_1)], \quad \forall s\in S. \tag{3.3}$$

Remark. The existence and uniqueness of the solution to (3.3) follows from a standard contraction argument (see Wang et al. [7]). However, existing literature has only shown the existence of a solution to (3.1) in the SA-rectangular settings under further assumptions [8, Theorem 8].

3.2 Bellman Optimality

Theorem 1. *If* (u^*,α^*) *solves* (3.1)*, then*

$$\begin{aligned}\alpha^* &= \overline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}}) = \underline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})\\ &= \overline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}}) = \underline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}}) \quad = \overline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{S}}) = \underline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{S}})\end{aligned} \tag{3.4}$$

for all $\mu\in\mathcal{P}(\mathcal{S})$*. Moreover, any other solution* (u,α) *to* (3.1) *satisfies* $\alpha=\alpha^*$.

Theorem 2. *If* (u^*,α^*) *solves* (3.1) *and* (3.2)*, then*

$$\sup_{\pi\in\Pi}\inf_{\kappa\in\mathrm{K}}\overline{\alpha}(\mu,\pi,\kappa) = \inf_{\kappa\in\mathrm{K}}\sup_{\pi\in\Pi}\overline{\alpha}(\mu,\pi,\kappa) = \sup_{\pi\in\Pi}\inf_{\kappa\in\mathrm{K}}\underline{\alpha}(\mu,\pi,\kappa) = \inf_{\kappa\in\mathrm{K}}\sup_{\pi\in\Pi}\underline{\alpha}(\mu,\pi,\kappa) = \alpha^*$$

for every combination of $\Pi=\Pi_{\mathrm{H}},\Pi_{\mathrm{S}}$ *and* $\mathrm{K}=\mathrm{K}_{\mathrm{H}},\mathrm{K}_{\mathrm{S}}$ *and any* $\mu\in\mathcal{P}(\mathcal{S})$.

4 Existence of Solution

4.1 General Criteria

Theorem 3. *Given arbitrary* $\mathcal{Q}\subset\mathcal{P}(\mathcal{A})$ *and* $\{\mathcal{P}_s\subset\{A\to\mathcal{P}(\mathcal{S})\}: s\in S\}$*, the following statements are equivalent:*

(1) The solutions $\{v_\gamma^*:\gamma\in(0,1)\}$ *to the* γ*-discounted equation* (3.3) *have uniformly bounded span; i.e.* $\sup_{\gamma\in(0,1)}|v_\gamma^*|_{\mathrm{span}} = \sup_{\gamma\in(0,1)}\left[\max_{s\in S} v_\gamma^*(s) - \min_{s\in S} v_\gamma^*(s)\right] < \infty$.

(2) The constant-gain average-reward robust Bellman equation (3.1) *has a solution* (u^*,α^*).

4.2 Communicating Structures

In this section, we show that under compactness and one-sided communicating structures, the Bellman equation does have solutions.

For $p \in \mathcal{P}$ and $\Delta : S \to \mathcal{P}(\mathcal{A})$, denote $p_\Delta(s'|s) := \sum_{a\in A} p(s'|s,a)\Delta(a|s)$. Also, let $p^n_\Delta(s'|s)$ be the (s,s') entry of the n'th power of the matrix $\{p_\Delta(s'|s) : s, s' \in S\}$.

Definition 3. Consider arbitrary controller and adversary action sets $\mathcal{Q} \subset \mathcal{P}(\mathcal{A})$ and $\mathcal{P} = \bigtimes_{s\in S} \mathcal{P}_s$, with $\mathcal{P}_s \subset \{A \to \mathcal{P}(\mathcal{S})\}$.

- A stationary controller policy $\Delta : S \to \mathcal{Q}$ is said to be communicating if for any $s, s' \in S$, there exists $p \in \mathcal{P}$ and $N \geq 1$ s.t. $p^N_\Delta(s'|s) > 0$. The controller is said to be communicating if all the stationary controller policies are communicating.
- A stationary adversary policy $p \in \mathcal{P}$ is said to be communicating if for any $s, s' \in S$, there exists $\Delta : S \to \mathcal{Q}$ and $N \geq 1$ s.t. $p^N_\Delta(s'|s) > 0$. The adversary is said to be communicating if all the stationary adversary policies are communicating.

Theorem 4. *If the controller is communicating and $\mathcal{Q}$ is compact, then the constant-gain average-reward robust Bellman equation* (3.1) *has a solution.*

Theorem 5. *If the adversary is communicating, $\mathcal{P}_s, s \in S$ are convex and compact, and $\mathcal{Q}$ is convex, then the constant-gain average-reward robust Bellman equation* (3.1) *has a solution.*

5 Bellman Optimality for the HD-S Case

In this section, we show a surprising result that, under a weak communication assumption, the average reward for a history-dependent controller against a stationary adversary corresponds to the solution of the Bellman equation with the inf-sup ordering, rather than its original form.

Definition 4. A stationary adversary policy $p \in \mathcal{P}$ is said to be weakly communicating if there is a communicating class C s.t. for any $s, s' \in C$, there exists $\Delta : S \to \mathcal{Q}$ and $N \geq 1$ s.t. $p^N_\Delta(s'|s) > 0$. Moreover, for all $s \in C^c$, s is transient under any stationary policy of the controller. The adversary is weakly communicating if every stationary policy $p \in \mathcal{P}$ is weakly communicating. Evidently, if the adversary is communicating, then it is also weakly communicating.

Proposition 5.1. *If $\{\delta_a : a \in A\} \subset \mathcal{Q}$ and the adversary is weakly communicating, then for each $\epsilon > 0$, there exists a history-dependent reinforcement learning policy $\pi_{\mathrm{RL}} \in \Pi_{\mathrm{H}}$ s.t.*

$$0 \leq \underline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{S}}) - \inf_{\kappa\in \mathrm{K}_{\mathrm{S}}} \underline{\alpha}(\mu, \pi_{\mathrm{RL}}, \kappa) \leq \inf_{\kappa\in \mathrm{K}_{\mathrm{S}}} \sup_{\pi\in\Pi_{\mathrm{S}}} \underline{\alpha}(\mu, \pi, \kappa) - \inf_{\kappa\in \mathrm{K}_{\mathrm{S}}} \underline{\alpha}(\mu, \pi_{\mathrm{RL}}, \kappa) \leq \epsilon$$

The same result holds true if $\underline{\alpha}$ is replaced with $\overline{\alpha}$.

Theorem 6. *If $\{\delta_a : a \in A\} \subset \mathcal{Q}$, then so long as the adversary is weakly communicating,*

$$\underline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{S}}) = \inf_{\kappa\in \mathrm{K}_{\mathrm{S}}} \sup_{\pi\in\Pi_{\mathrm{H}}} \underline{\alpha}(\mu, \pi, \kappa) = \inf_{\kappa\in \mathrm{K}_{\mathrm{S}}} \sup_{\pi\in\Pi_{\mathrm{S}}} \underline{\alpha}(\mu, \pi, \kappa).$$

The same result holds true if $\underline{\alpha}$ is replaced with $\overline{\alpha}$. Moreover, if (3.2) *admits a solution pair (u', α'), then $\underline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{S}}) = \overline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{S}}) = \alpha'$. A sufficient condition for* (3.2) *to have a solution is that the adversary is communicating and $\mathcal{P}_s : s \in S$ are compact.*

Intuitively, when the stationary adversary is weakly communicating, a history-dependent controller policy can adaptively "learn" the adversary policy through online reinforcement learning. Importantly, this learning process doesn't affect the long-run average performance of the controller policy, hence achieves the inf-sup value.

In particular, this implies that if solutions exist for (3.1) and (3.2), but the corresponding gains α^* and α' do not coincide, then stationary optimality cannot be expected for the robust optimal control problems $\underline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{S}})$ and $\overline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{S}})$. A converse of this is also true. This is summarized in the following Corollary 6.1.

Corollary 6.1. *Assume that both the controller and the adversary are communicating, $\{\delta_a : a \in A\} \subset \mathcal{Q}$, and that $\mathcal{Q}$ and $\mathcal{P}_s,\ s \in S$ are compact. If the adversary's policy class is stationary, i.e., $\mathrm{K} = \mathrm{K}_{\mathrm{S}}$, then stationary policies are optimal for a history-dependent controller $\Pi = \Pi_{\mathrm{H}}$ if and only if $\alpha' = \alpha^*$.*

References

[1] Bartlett, P. L. and Tewari, A. (2009). Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 35–42, Arlington, Virginia, USA. AUAI Press.

[2] Berge, C. (1963). *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces and Convexity*. Oliver & Boyd.

[3] Filar, J. and Vrieze, K. (2012). *Competitive Markov decision processes*. Springer Science & Business Media.

[4] Grand-Clement, J., Petrik, M., and Vieille, N. (2023). Beyond discounted returns: Robust markov decision processes with average and blackwell optimality. *arXiv preprint arXiv:2312.03618*.

[5] Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280.

[6] Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

[7] Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023a). On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*.

[8] Wang, Y., Velasquez, A., Atia, G., Prater-Bennette, A., and Zou, S. (2023b). Robust average-reward markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[9] Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.

[10] Xu, H. and Mannor, S. (2010). Distributionally robust markov decision processes. In *NIPS*, pages 2505–2513.

[11] Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR.

Appendices

A Proofs for Section 3

Recall that the history

$$\mathbf{H}_t := \{h_t = (s_0, a_0, \dots, a_{t-1}, s_t) : \omega = (s_0, a_0, \dots, a_{t-1}, s_t, \dots) \in \Omega\}.$$

We also define the random element $H_t : \Omega \to \mathbf{H}_t$ by point evaluation $H_t(\omega) = h_t$, and the σ-field $\mathcal{H}_t := \sigma(H_t)$. Next, we define $\{\mathbf{G}_t : t \geq 0\}$ by

$$\mathbf{G}_t := \{g_t = (s_0, a_0, \dots, s_t, a_t) : \omega = (s_0, a_0, \dots, s_t, a_t \dots) \in \Omega\}.$$

Note that g_t is the concatenation of the history h_t with the controller's action at time t, i.e., $g_t = (h_t, a_t)$, where $h_t \in \mathbf{H}_t$. Also, define the random element $G_t : \Omega \to \mathbf{G}_t$ by point evaluation $G_t(\omega) = g_t$, and $\mathcal{G}_t := \sigma(G_t)$.

To prove the main theorems in Section 3, we introduce an important technical tool.

Proposition A.1. *For any function $f : S \to \mathbb{R}$ and any pair of policies $\pi \in \Pi_{\mathrm{H}}$ and $\kappa \in \mathrm{K}_{\mathrm{H}}$, define the process*

$$M^{\pi,\kappa}_{f,n} = \sum_{k=1}^{n} f(X_k) - \sum_{a,s'} \pi_{k-1}(a|H_{k-1})\kappa_{k-1}(s'|H_{k-1}, a) f(s').$$

Then, $M^{\pi,\kappa}_{f,n}$ is a $\mathcal{H}_k, P^{\pi,\kappa}_{\mu}$-Martingale.

A.1 Proof of Proposition A.1

Proof. It suffices to check $E[M^{\pi,\kappa}_{f,k} - M^{\pi,\kappa}_{f,k-1}|\mathcal{H}_{k-1}] = 0$.

We see that the conditional distribution of (A_{k-1}, X_k) given H_{k-1} is determined by π_{k-1} and κ_{k-1}. So,

$$E^{\pi,\kappa}_{\mu}\left[f(X_k)|\mathcal{H}_{k-1}\right] = \sum_{a,s'} \pi_{k-1}(a|H_{k-1})\kappa_{k-1}(s'|H_{k-1}, a) f(s').$$

Also, note that

$$M^{\pi,\kappa}_{f,k} - M^{\pi,\kappa}_{f,k-1} = f(X_k) - \sum_{a,s'} \pi_{k-1}(a|H_{k-1})\kappa_{k-1}(s'|H_{k-1}, a) f(s').$$

This completes the proof. □

A.2 Proof of Theorem 1

Proof. Note that the second claim follows from the first claim: If (u, α) is any other solution to (3.1), then by (3.4), $\alpha = \overline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{H}})$. On the other hand, by (3.4), $\overline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{H}}) = \alpha^*$. So, $\alpha = \alpha^*$.

To show (3.4), observe that $\underline{\alpha}(\mu, \Pi_{\mathrm{S}}, \mathrm{K}_{\mathrm{H}})$ is the smallest maxmin control average-reward among the ones that appear in (3.4). We will first show that $\underline{\alpha}(\mu, \Pi_{\mathrm{S}}, \mathrm{K}_{\mathrm{H}}) \geq \alpha^*$. Then, we show $\overline{\alpha}(\mu, \Pi_{\mathrm{S}}, \mathrm{K}_{\mathrm{S}}) \leq \alpha^*$ as well as $\overline{\alpha}(\mu, \Pi_{\mathrm{H}}, \mathrm{K}_{\mathrm{H}}) \leq \alpha^*$. Combining these, we can conclude (3.4) and hence Theorem 1.

Step 1: Show $\underline{\alpha}(\mu, \Pi_{\mathrm{S}}, \mathrm{K}_{\mathrm{H}}) \geq \alpha^*$.

Since (u^*, α^*) solves (3.1), for each $\epsilon > 0$, there exists a controller decision rule $\Delta : S \to \mathcal{P}(\mathcal{A})$ so that

$$\inf_{p_s \in \mathcal{P}_s} E_{\Delta(\cdot|s), p_s}[r(s, A_0) - \alpha^* + u(X_1)] \geq u^*(s) - \epsilon.$$

Therefore, for any history-dependent adversarial policy $\kappa = (\kappa_0, \kappa_1, \dots)$ and any $s \in S, g_{k-1} \in \mathbf{G}_{k-1}, k \geq 0$, we have that

$$\sum_{a \in A} \Delta(a|s) \left(r(s, a) - \alpha + \sum_{s' \in S} \kappa_k(s'|g_{k-1}, s, a) u^*(s') \right) \geq u^*(s) - \epsilon. \tag{A.1}$$

Using (A.1), we have that

$$\begin{aligned}
&\sum_{k=0}^{n-1}\sum_{a\in A}\Delta(a|X_t)[r(X_t,a)-\alpha^*] \\
&\geq -n\epsilon + \sum_{k=0}^{n-1}\left[u^*(X_k) - \sum_{(a,s')\in A\times S}\kappa_k(s'|H_k,a)\Delta(a|X_k)u^*(s')\right] \\
&= -n\epsilon + M_{u^*,n}^{\Delta,\kappa} - u^*(X_n) + u^*(X_0)
\end{aligned} \tag{A.2}$$

On the other hand, notice that

$$E_\mu^{\Delta,\kappa}[r(X_k,A_k)-\alpha^*] = E_\mu^{\Delta,\kappa}\sum_{a\in A}\Delta(a|X_k)[r(X_k,a)-\alpha^*]. \tag{A.3}$$

Therefore, by (A.2),

$$\begin{aligned}
\frac{1}{n}\sum_{k=0}^{n-1}E_\mu^{\Delta,\kappa}[r(X_k,A_k)-\alpha^*] &= \frac{1}{n}E_\mu^{\Delta,\kappa}\sum_{k=0}^{n-1}\sum_{a\in A}\Delta(a|X_k)[r(X_k,a)-\alpha^*] \\
&\geq -\epsilon + E_\mu^{\Delta,\kappa}M_{u^*,n}^{\Delta,\kappa} + E_\mu^{\Delta,\kappa}\frac{u^*(X_0)-u^*(X_n)}{n} \\
&\to -\epsilon
\end{aligned}$$

as $n\to\infty$. Here, we use Proposition A.1 to conclude that $E_\mu^{\Delta,\kappa}M_{u^*,n}^{\Delta,\kappa}=0$

So, we have that for arbitrary $\kappa\in \mathrm{K_H}$,

$$\liminf_{n\to\infty} E_\mu^{\Delta,\kappa}\frac{1}{n}\sum_{k=0}^{n-1}r(X_k,A_k) \geq \alpha^* - \epsilon.$$

This implies that

$$\inf_{\kappa\in \mathrm{K_H}}\liminf_{n\to\infty} E_\mu^{\Delta,\kappa}\frac{1}{n}\sum_{k=0}^{n-1}r(X_k,A_k) \geq \alpha^* - \epsilon.$$

Moreover, since $\Delta\in\Pi_\mathrm{S}$, we have that

$$\underline{\alpha}(\mu,\Pi_\mathrm{S},\mathrm{K_H}) = \sup_{\pi\in\Pi_\mathrm{S}}\inf_{\kappa\in\mathrm{K_H}}\underline{\alpha}(\mu,\pi,\kappa) \geq \alpha^*-\epsilon. \tag{A.4}$$

Since $\epsilon>0$ is arbitrary, we conclude that $\underline{\alpha}(\mu,\Pi_\mathrm{S},\mathrm{K_H})\geq\alpha^*$.

Step 2: Show $\overline{\alpha}(\mu,\Pi_\mathrm{S},\mathrm{K_S})\leq\alpha^*$ and $\overline{\alpha}(\mu,\Pi_\mathrm{H},\mathrm{K_H})\leq\alpha^*$.

We consider an arbitrary history-dependent policy $\pi=(\pi_0,\pi_1,\dots)\in\Pi_\mathrm{H}$. Since (u^*,α^*) solves (3.1), for any $s\in S$, $g_{k-1}\in\mathbf{G}_{k-1}$ and $k\geq 0$,

$$\inf_{p_s\in\mathcal{P}_s}E_{\pi_k(\cdot|g_{k-1},s),p_s}[r(s,A_0)-\alpha^*+u^*(X_1)]\leq u^*(s).$$

Hence, there exists $\kappa_k(\cdot|g_{k-1},s,\cdot)\in\mathcal{P}_s$ so that

$$\sum_{a\in A}\pi_k(a|g_{k-1},s)\left(r(s,a)+\sum_{s'\in S}\kappa_k(s'|g_{k-1},s,a)u^*(s')\right)\leq u^*(s)+\epsilon \tag{A.5}$$

for each $s\in S$, $g_{k-1}\in\mathbf{G}_{k-1}$ and $k\geq 0$.

Moreover, by the same argument, if $\pi_k(\cdot|g_{k-1},s)=\pi(\cdot|s)$ is Markov time-homogeneous, there exists $\kappa_k(\cdot|g_{k-1},s,\cdot)=\kappa(\cdot|s,\cdot)$ Markov time-homogeneous so that (A.5) holds. So, we have constructed an adversarial policy $\kappa:=(\kappa_0,\kappa_1,\dots)\in\mathrm{K_H}$ (or $\mathrm{K_S}$) with $\{\kappa_k:k\geq 0\}$ specified above.

Therefore, with (A.5), we have that

$$\begin{aligned}
&\sum_{k=0}^{n-1}\sum_{a\in A}\pi_k(a|H_k)[r(X_k,a)-\alpha^*]\\
&\le n\epsilon+\sum_{k=0}^{n-1}\left[u^*(X_k)-\sum_{(a,s')\in A\times S}\pi_k(a|H_k)\kappa_k(s'|H_k,a)u^*(s')\right]\\
&= n\epsilon+M^{\pi,\kappa}_{u^*,n}-u^*(X_n)+u^*(X_0)
\end{aligned}\tag{A.6}$$

As in (A.3), notice that

$$E^{\pi,\kappa}_\mu[r(X_k,A_k)-\alpha^*]=E^{\pi,\kappa}_\mu\sum_{a\in A}\pi_k(a|H_k)[r(X_k,a)-\alpha^*].$$

Therefore, by (A.6),

$$\begin{aligned}
\frac{1}{n}\sum_{k=0}^{n-1}E^{\pi,\kappa}_\mu[r(X_k,A_k)-\alpha^*] &= \frac{1}{n}E^{\pi,\kappa}_\mu\sum_{k=0}^{n-1}\sum_{a\in A}\pi_k(a|H_k)[r(X_k,a)-\alpha^*]\\
&\le \epsilon+E^{\pi,\kappa}_\mu M^{\pi,\kappa}_{u^*,n}+E^{\pi,\kappa}_\mu\frac{u^*(X_0)-u^*(X_n)}{n}\\
&\to\epsilon
\end{aligned}$$

as $n\to\infty$. Here, we also use Proposition A.1 to conclude that $E^{\pi,\kappa}_\mu M^{\pi,\kappa}_{u^*,n}=0$

So, we have that that for arbitrary $\pi\in\Pi_{\mathrm{H}}$ (or $\pi\in\Pi_{\mathrm{S}}$) there exists $\kappa\in\mathrm{K}_{\mathrm{H}}$ (or $\kappa\in\mathrm{K}_{\mathrm{S}}$) s.t.

$$\limsup_{n\to\infty}E^{\pi,\kappa}_\mu\frac{1}{n}\sum_{k=0}^{n-1}r(X_k,A_k)\le\alpha^*+\epsilon.$$

Hence,

$$\inf_{\kappa\in\mathrm{K}_{\mathrm{H}}}\limsup_{n\to\infty}E^{\pi,\kappa}_\mu\frac{1}{n}\sum_{k=0}^{n-1}r(X_k,A_k)\le\alpha^*-\epsilon$$

where $\kappa\in\mathrm{K}_{\mathrm{H}}$ is replaced by $\kappa\in\mathrm{K}_{\mathrm{S}}$ if $\pi\in\Pi_{\mathrm{S}}$.

Moreover, since $\pi\in\Pi_{\mathrm{H}}$ (or $\pi\in\Pi_{\mathrm{S}}$) can be any policy, we have that

$$\overline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})=\sup_{\pi\in\Pi_{\mathrm{H}}}\inf_{\kappa\in\mathrm{K}_{\mathrm{H}}}\overline{\alpha}(\mu,\pi,\kappa)\le\alpha^*-\epsilon.\tag{A.7}$$

Since $\epsilon>0$ can be arbitrarily small, we conclude that $\overline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})\le\alpha^*$. The same proof still holds when $\pi\in\Pi_{\mathrm{S}}$ and $\kappa\in\mathrm{K}_{\mathrm{S}}$, leading to $\overline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{S}})\le\alpha^*$.

Step 3: Combining Steps 1 and 2 We combine the results from steps 1 and 2 to conclude that

$$\begin{aligned}
\alpha^*&\le\underline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}})\le\underline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})\le\overline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})\le\alpha^*,\\
\alpha^*&\le\underline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}})\le\overline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}})\le\overline{\alpha}(\mu,\Pi_{\mathrm{H}},\mathrm{K}_{\mathrm{H}})\le\alpha^*,\\
\alpha^*&\le\underline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}})\le\underline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{S}})\le\overline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{S}})\le\alpha^*,\\
\alpha^*&\le\underline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}})\le\overline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{H}})\le\overline{\alpha}(\mu,\Pi_{\mathrm{S}},\mathrm{K}_{\mathrm{S}})\le\alpha^*.
\end{aligned}$$

These inequalities imply (3.4). □

A.3 Proof of Theorem 2

Proof. Since (u^*,α^*) solves (3.2), we have that there exists $\psi: S\times A\to\mathcal{P}(\mathcal{S})$ s.t. $\psi(\cdot|s,\cdot)\in\mathcal{P}_s$ for all $s\in S$ and

$$\sup_{\phi\in\mathcal{Q}}E_{\phi,\psi(\cdot|s,\cdot)}[r(s,A_0)-\alpha^*+u(X_1)]\le u^*(s)+\epsilon.$$

Thus, for any history-dependent policy $\pi = (\pi_0, \pi_1, \dots)$, and $s \in S, g_{k-1} \in \mathbf{G}_{k-1}, k \geq 0$,

$$\sum_{a\in A} \pi_k(a|g_{k-1}, s) \left[r(s,a) - \alpha^* + \sum_{s'\in S} \psi(s'|s,a) u^*(s') \right] \leq u^*(s) + \epsilon. \tag{A.8}$$

Note that

$$\begin{aligned}
\frac{1}{n}\sum_{k=0}^{n-1} E_\mu^{\pi,\psi}[r(X_k, A_k) - \alpha^*] &= \frac{1}{n}\sum_{k=0}^{n-1} E_\mu^{\pi,\psi} E_\mu^{\pi,\psi}[r(X_k, A_k) - \alpha^* | \mathcal{H}_k] \\
&= \frac{1}{n}\sum_{k=0}^{n-1} E_\mu^{\pi,\psi} \left[\sum_a \pi_k(a|G_{k-1}, X_k) r(X_k, a) - \alpha^* \right] \\
&\overset{(i)}{\leq} \frac{1}{n}\sum_{k=0}^{n-1} \left(E_\mu^{\pi,\psi}\left[u^*(X_k) + \epsilon\right] - E_\mu^{\pi,\psi}\left[\sum_{a\in A, s'\in S} \pi_k(a|H_k)\psi(s'|s,a)u^*(s') \right] \right) \\
&= \epsilon + \frac{1}{n} E_\mu^{\pi,\psi} M_n^{\pi,\psi} + E_\mu^{\pi,\psi} \frac{u^*(X_0) - u^*(X_n)}{n} \\
&= \epsilon + E_\mu^{\pi,\psi} \frac{u^*(X_0) - u^*(X_n)}{n} \\
&\to \epsilon
\end{aligned}$$

$n \to \infty$, where (i) follows from (A.8).

So, we have that

$$\limsup_{n\to\infty} E_\mu^{\pi,\psi} \frac{1}{n} \sum_{k=0}^{n-1} r(X_k, A_k) \leq \alpha^* + \epsilon.$$

Since $\pi \in \Pi_\mathrm{H}$ is arbitrary, we can conclude that

$$\inf_{\kappa\in\mathrm{K_S}} \sup_{\pi\in\Pi_\mathrm{H}} \overline{\alpha}(\mu,\pi,\kappa) \leq \sup_{\pi\in\Pi_\mathrm{H}} \overline{\alpha}(\mu,\pi,\psi) \leq \alpha^* + \epsilon. \tag{A.9}$$

On the other hand, since (u^*, α^*) solves (3.1), Theorem 1 and the proofs in Appendix A.2 are still valid. In particular, by (A.4), still holds. Therefore, combining (A.4) with (A.9), we have that

$$\begin{aligned}
\alpha^* - \epsilon &\leq \sup_{\pi\in\Pi_\mathrm{S}} \inf_{\kappa\in\mathrm{K_H}} \underline{\alpha}(\mu,\pi,\kappa) \\
&\leq \sup_{\pi\in\Pi_\mathrm{H}} \inf_{\kappa\in\mathrm{K_H}} \underline{\alpha}(\mu,\pi,\kappa) \\
&\leq \inf_{\kappa\in\mathrm{K_H}} \sup_{\pi\in\Pi_\mathrm{H}} \underline{\alpha}(\mu,\pi,\kappa) \\
&\leq \inf_{\kappa\in\mathrm{K_S}} \sup_{\pi\in\Pi_\mathrm{H}} \overline{\alpha}(\mu,\pi,\kappa) \\
&\leq \alpha^* + \epsilon.
\end{aligned}$$

Since ϵ is arbitrary, we have that

$$\alpha^* = \sup_{\pi\in\Pi_\mathrm{S}} \inf_{\kappa\in\mathrm{K_H}} \underline{\alpha}(\mu,\pi,\kappa) = \inf_{\kappa\in\mathrm{K_S}} \sup_{\pi\in\Pi_\mathrm{H}} \overline{\alpha}(\mu,\pi,\kappa).$$

This implies the statement of the theorem, as $\sup_{\pi\in\Pi_\mathrm{S}} \inf_{\kappa\in\mathrm{K_H}} \underline{\alpha}(\mu,\pi,\kappa)$ is the smallest and $\inf_{\kappa\in\mathrm{K_S}} \sup_{\pi\in\Pi_\mathrm{H}} \overline{\alpha}(\mu,\pi,\kappa)$ is the largest among all the relevant values in the statement of the theorem. □

B Proofs for Section 4

B.1 Proof of Theorem 3

Proof. (1) $\Longrightarrow$ (2):

We fix a reference state $s_0 \in S$ and define

$$u_\gamma = v_\gamma^* - v_\gamma^*(s_0), \qquad \alpha_\gamma = (1-\gamma)v_\gamma^*(s_0).$$

Since v_γ^* solves (3.3), we observe that

$$\begin{aligned} v_\gamma^*(s) - v_\gamma^*(s_0) &= \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) + \gamma(v_\gamma^*(X_1) - v_\gamma^*(s_0))] - (1-\gamma)v_\gamma^*(s_0) \\ u_\gamma(s) &= \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha_\gamma + \gamma u_\gamma(X_1)] \end{aligned} \tag{B.1}$$

From Wang et al. [7], $\|v_\gamma^*\|_\infty \le 1/(1-\gamma)$. So, $0 \le \alpha_\gamma \le 1$. Moreover, by (1), $\|u_\gamma\|_\infty \le |v_\gamma^*|_{\text{span}} \le C < \infty$ uniformly in γ. Hence $(u_\gamma, \alpha_\gamma) \in [-C, C]^{|S|} \times [0,1]$ for all γ. As $[-C, C]^{|S|} \times [0,1]$ is compact in the sup metric, there exists a convergent subsequence $\{(u_{\gamma_n}, \alpha_{\gamma_n}) : n = 1, 2, \ldots\}$ with $(u_*, \alpha_*) := \lim_{n\to\infty}(u_{\gamma_n}, \alpha_{\gamma_n})$.

Next, we would like to take the limit $n \to \infty$ on both sides of (B.1), with γ replaced by γ_n. To do this, we define for $\gamma \in [0,1]$, $(u,\alpha) \in [-C,C]^{|S|} \times [0,1]$, $\phi \in \mathcal{P}(\mathcal{A})$, $p_s \in \{A \to \mathcal{P}(\mathcal{S})\}$,

$$f_s(\gamma, u, \alpha, \phi, p_s) = E_{\phi,p_s}[r(s,A_0) - \alpha + \gamma u(X_1)],$$

which is a continuous function.

We first note that since $\mathcal{P}_s$ is bounded and the mapping $p_s \to f_s(\gamma,u,\alpha,\phi,p_s)$ is continuous,

$$\inf_{p_s\in\mathcal{P}_s} f_s(\gamma,u,\alpha,\phi,p_s) = \min_{p_s\in\overline{\mathcal{P}}_s} f_s(\gamma,u,\alpha,\phi,p_s)$$

where $\overline{\mathcal{P}}_s$ is the closure of $\mathcal{P}_s$.

Since $\overline{\mathcal{P}}_s$ is compact and does not depend on γ, ϕ, u, α, by Berge's maximum theorem [2, VI.3, Theorem 1 & 2], the mapping

$$(\gamma,u,\alpha,\phi) \to m_s(\gamma,u,\alpha,\phi) := \min_{p_s\in\overline{\mathcal{P}}_s} f_s(\gamma,u,\alpha,\phi,p_s)$$

is continuous for $\gamma \in [0,1]$, $(u,\alpha) \in [-C,C]^{|S|} \times [0,1]$, and $\phi \in \mathcal{P}(\mathcal{A})$.

Apply the same argument, we have that

$$M_s(\gamma,u,\alpha) = \max_{\phi\in\overline{\mathcal{Q}}} m_s(\gamma,u,\alpha,\phi) = \sup_{\phi\in\mathcal{Q}} m_s(\gamma,u,\alpha,\phi) = \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} f_s(\gamma,u,\alpha,\phi,p_s)$$

is continuous for $\gamma \in [0,1]$ and $(u,\alpha) \in [-C,C]^{|S|} \times [0,1]$.

Therefore, we have that

$$\lim_{n\to\infty} M_s(\gamma_n, u_{\gamma_n}, \alpha_{\gamma_n}) = M_s(1, u_*, \alpha_*) = \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha_* + u_*(X_1)].$$

This and (B.1) implies that

$$u_*(s) = \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha_* + u_*(X_1)]$$

i.e. (u_*, α_*) solves (3.1).

(2) $\implies$ (1):

Let (u^*, α^*) be a solution to (3.1). Due to the solutions of (3.1) being shift-invariant, w.l.o.g., we assume that $u^* \ge 0$ and $\min_{s\in S} u(s) = 0$. To simplify notation, we define the discounted Bellman operator

$$\mathcal{T}_\gamma[v] := \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) + \gamma v(X_1)].$$

Then $\mathcal{T}_\gamma[v_\gamma^*] = v_\gamma^*$, where v_γ^* is the unique fixed-point.

We define two auxiliary values

$$\overline{v}_\gamma := \frac{\alpha^*}{1-\gamma} + u^*, \quad \underline{v}_\gamma := \frac{\alpha^*}{1-\gamma} + u^* - |u^*|_{\text{span}}. \tag{B.2}$$

Step 1: We show that $\mathcal{T}_\gamma[\overline{v}_\gamma] \leq \overline{v}_\gamma$ and $\mathcal{T}_\gamma[\underline{v}_\gamma] \geq \underline{v}_\gamma$.

We observe that for all $s \in S$,

$$\begin{aligned}
\mathcal{T}_\gamma[\overline{v}_\gamma](s) &= \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) + \gamma\overline{v}_\gamma(X_1)] \\
&= \frac{\gamma\alpha^*}{1-\gamma} + \alpha^* + \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha^* + \gamma u^*(X_1)] \\
&= \frac{\alpha^*}{1-\gamma} + \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha^* + u^*(X_1) - (1-\gamma)u^*(X_1)] \\
&\overset{(i)}{\leq} \frac{\alpha^*}{1-\gamma} + \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha^* + \gamma u^*(X_1)] \\
&= \overline{v}_\gamma(s)
\end{aligned}$$

where (i) follows from the choice that $u^* \geq 0$ and the last equality uses the assumption that (u^*, α^*) solves (3.1). On the other hand,

$$\begin{aligned}
\mathcal{T}_\gamma[\underline{v}_\gamma](s) &= \frac{\alpha^*}{1-\gamma} + \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha^* + \gamma(u^*(X_1) - |u^*|_{\text{span}})] \\
&= \frac{\alpha^*}{1-\gamma} - |u^*|_{\text{span}} + \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha^* + u^*(X_1) + (1-\gamma)(|u^*|_{\text{span}} - u^*(X_1))] \\
&\overset{(i)}{\geq} \frac{\alpha^*}{1-\gamma} - |u^*|_{\text{span}} + \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) - \alpha^* + u^*(X_1)] \\
&= \underline{v}_\gamma(s)
\end{aligned}$$

where (i) follows from $u^* \geq 0$ and $\min_{s\in S} u^*(s) = 0$ hence $|u^*|_{\text{span}} - u^* = \|u^*\|_\infty - u^* \geq 0$.

Step 2: We prove that v^*_γ, the solution to (3.3), is upper bounded by $\overline{v}_\gamma$ and lower bounded by $\underline{v}_\gamma$; i.e.

$$\underline{v}_\gamma \leq v^*_\gamma \leq \overline{v}_\gamma.$$

To achieve this, we will use the fact that $\mathcal{T}_\gamma$ is a monotone γ-contraction.

First, the contraction property of $\mathcal{T}_\gamma$ is well known (see Wang et al. [7]). We then show that $\mathcal{T}_\gamma$ is a monotone operator; i.e., $\mathcal{T}_\gamma[u] \geq \mathcal{T}_\gamma[v]$ if $u \geq v$. This is straightforward

$$\mathcal{T}_\gamma[u](s) = \sup_{\phi\in\mathcal{Q}} \inf_{p_s\in\mathcal{P}_s} E_{\phi,p_s}[r(s,A_0) + \gamma[u(X_1) - v(X_1)] + \gamma v(X_1)] \geq \mathcal{T}_\gamma[v](s)$$

where we used that $u(X_1) - v(X_1) \geq 0$.

Next, we check by induction that

$$\mathcal{T}^k_\gamma[\overline{v}_\gamma] := \underbrace{(\mathcal{T}_\gamma \circ \cdots \circ \mathcal{T}_\gamma)}_{\times k}[\overline{v}] \leq \overline{v}$$

for all $k \geq 1$. The base case $k = 1$ follows from the previous proof. For the induction step, assume that $\mathcal{T}^k_\gamma[\overline{v}_\gamma] \leq \overline{v}_\gamma$. By the monotonicity of $\mathcal{T}_\gamma$, we have that

$$\mathcal{T}^{k+1}_\gamma[\overline{v}_\gamma] = \mathcal{T}_\gamma[\mathcal{T}^k_\gamma[\overline{v}_\gamma]] \leq \mathcal{T}_\gamma[\overline{v}_\gamma] \leq \overline{v}_\gamma,$$

completing the induction step.

On the other hand, by the contraction property, $\overline{v}_\gamma \geq \mathcal{T}^k_\gamma[\overline{v}_\gamma] \to v^*_\gamma$ as $k \to \infty$. So, we have that $\overline{v}_\gamma \geq v^*_\gamma$.

Similarly, we show that v^*_γ is lower bounded by $\underline{v}_\gamma$. Again, we apply the same induction argument. We see that the base case holds due to $\mathcal{T}_\gamma[\underline{v}_\gamma] \geq \underline{v}_\gamma$, and the induction step follows from the monotonicity of $\mathcal{T}_\gamma$. Therefore, by the contraction property, $\underline{v}_\gamma \leq \mathcal{T}^k_\gamma[\underline{v}_\gamma] \to v^*_\gamma$ as $k \to \infty$. So, we have that $\underline{v}_\gamma \leq v^*_\gamma$.

Step 3: We conclude the proof by bounding the span of v_γ^*.

Since $\underline{v}_\gamma \leq v_\gamma^* \leq \overline{v}_\gamma$,

$$\left|v_\gamma^*\right|_{\text{span}} = \max_{s\in S} v_\gamma^*(s) - \min_{s\in S} v_\gamma^*(s) \leq \max_{s\in S} \overline{v}_\gamma(s) - \min_{s\in S} \underline{v}_\gamma(s) = 2\left|u^*\right|_{\text{span}}$$

where the last equality follows from the definition of $\overline{v}_\gamma$ and $\underline{v}_\gamma$ in (B.2). $\square$

B.2 Proof of Theorem 4

Proof. By Theorem 3, it suffices to show that under the assumptions of Theorem 4, the solution v_γ^* to (3.3) has uniformly bounded span. To show this, we take the following steps.

Step 1: We will construct a finite subset of controller's policy C that ensures a uniform communication probability lower bound. We use this property to show an enhanced version of the communicating controller under compact $\mathcal{Q}$ in Lemma 1.

We consider an arbitrary fixed pair of states $x, y \in S$. Since the controller is assumed to be communicating, for any stationary policy $\Delta : S \to \mathcal{Q}$, there exists $p = p^{x,y,\Delta} \in \mathcal{P}$ and $N = N^{x,y,\Delta} \geq 1$ s.t. $p_\Delta^N(y|x) > 0$. Note that both p and N are dependent on Δ, x, y. We will suppress the dependence for notation simplicity.

Note that if we fix $p = p^\Delta$ and $N = N^\Delta$, then the mapping $\eta \to p_\eta^N$ for $\eta : S \to \mathcal{P}(\mathcal{A})$ is continuous in η. This is because both $\eta \to p_\eta$ and $p_\eta \to p_\eta^N$ are continuous. So, there must exists an open neighborhood G_Δ of Δ s.t. $p_\eta^N(y|x) \geq p_\Delta^N(y|x)/2$ for all $\eta \in G_\Delta$.

Since $\mathcal{Q}$ is compact, $\{S \to \mathcal{Q}\}$, seen as stochastic matrices in $\mathbb{R}^{|S|\times|A|}$, is also compact. Note that clearly, $\{G_\Delta : \Delta \in \{S \to \mathcal{Q}\}\}$ is an open cover of $\{S \to \mathcal{Q}\}$. Hence, by the compactness of $\{S \to \mathcal{Q}\}$, there exists a finite sub-cover $\{G_\Delta : \Delta \in C\}$ where $C := \left\{\Delta_1, \ldots, \Delta_{|C|}\right\} \subset \{S \to \mathcal{Q}\}$ is a finite subset.

Since C is finite, by the construction of G_Δ, we have that

$$\min_{\Delta\in C} (p_\Delta^\Delta)^{N^\Delta}(y|x) =: \delta(y|x) > 0.$$

Note that $\delta(y|x)$ is independent of Δ, and here the dependence of N and p on x, y are suppressed. On the other hand, for any stationary policy η, $\eta \in G_{\Delta_k}$ for some $k \in \{1, \ldots, |C|\}$. So,

$$(p_\eta^{\Delta_k})^{N^{\Delta_k}}(y|x) \geq \frac{1}{2}(p_{\Delta_k}^{\Delta_k})^{N^{\Delta_k}}(y|x) \geq \delta(y|x) > 0.$$

Since the state space is finite, we can define

$$\delta := \min_{x,y\in S} \delta(y|x), \qquad M := \max_{k=1,\ldots,|C|;\, x,y\in S} N^{x,y,\Delta_k}.$$

This implies the following Lemma 1.

Lemma 1. *Under the assumptions of Theorem 4, there exists $\delta > 0$ and positive integer M s.t. for any stationary controller policy $\Delta : S \to \mathcal{Q}$ and any pair of states $x, y \in S$, there exists $p \in \mathcal{P}$ and $N \leq M$ s.t. $p_\Delta^N(y|x) \geq \delta$.*

Step 2: We show that under Lemma 1, for fixed $y \in S$ and policy $\Delta : S \to \mathcal{Q}$, there exists a $p \in \mathcal{P}$ independent of the initial state $x \in S$ so that the probability of the first hitting time $\tau_y := \inf\{k \geq 0 : X_k = y\}$ being less than or equal to M has a uniform lower bound. This is summarized in Lemma 2.

Lemma 2. *Under the assumptions of Theorem 4, there exists $\delta' > 0$ s.t. for any stationary controller policy $\Delta : S \to \mathcal{Q}$ and states $y \in S$, there exists $q \in \mathcal{P}$ s.t.*

$$\min_{x\in S} P_x^{\Delta,q}(\tau_y \leq |S|) \geq \delta'.$$

We fix $y \in S$ and $\Delta : S \to \mathcal{Q}$. To verify Lemma 2, note that in Lemma 1 we can take $M \leq |S|$, since the longest loop-free path between two states in S has at most $|S|$ steps.

For each $x \in S$, let $p = p^{x,y,\Delta}$ the adversarial stationary policy in Lemma 1. Then, we have that

$$P_x^{\Delta,p}(\tau_y \leq M) \geq p_\Delta^N(y|x) \geq \delta.$$

So, to prove Lemma 2, we can show that the choice $p^{x,y,\Delta}$ can be made independent of x.

Let us define $q = q^{y,\Delta} \in \mathcal{P}$ algorithmically as follows. We will iteratively assign $q(\cdot|s,\cdot) \in \mathcal{P}_s$ until all $\{q(\cdot|s,\cdot) : s \in S\}$ has been assigned. Denote $q_\Delta(\cdot|s) = \sum_{a\in A} \Delta(a|s_i)q(\cdot|s,a)$.

We initialize the algorithm by assigning $q(\cdot|y,\cdot) = p_y$ for an arbitrary $p_y \in \mathcal{P}_y$. Then, Let $V = \{y\}$ be the assigned states, and V^c the compliment in S are the unassigned states.

1. Choose any unassigned state $s_0 \in V^c$. Then by Lemma 1 there exists $p = p^{s_0,y,\Delta} \in \mathcal{P}$ and N s.t. $p_\Delta^N(y|s_0) \geq \delta$. Therefore, there exists a path $s_0 \to s_1 \to, \ldots, \to s_N = y$ s.t. $p_\Delta(s_{k+1}|s_k) > 0$.

   Moreover, since there are at most $P_N := (|S|-1)!/(|S|-N)!$ paths from s_0 to y in N steps, there must be one path with probability at least δ/P_N under p_Δ. Let $s_0 \to s_1 \to, \ldots, \to s_N = y$ be this path.

   Note that, in general, this path could be repeating, i.e., $s_i = s_j$ for some $i < j$. However, we can "trim off" the in-between segment to get $s_0 \to \ldots \to s_i \to s_{j+1} \to \ldots \to s_N = y$. This is again a path with probability at least δ/P_N under p_Δ. We trim until obtaining a non-repeating path with probability at least δ/P_N and relabel it with $s_0 \to s_1 \to, \ldots, \to s_k = y$ for some $k \leq N$.

   Therefore, we have that on this path, for all $i \leq k-1$,

   $$p_\Delta(s_{i+1}|s_i) \geq \prod_{i=0}^{k-1} p_\Delta(s_{i+1}|s_i) \geq \delta/P_N$$

   where δ/P_N is independent of x, y, Δ.

2. Let $j = \min\{i \geq 1 : s_i \in V\}$ be the first index so that s_j is assigned. So, $s_i \in V^c$ for all $i \leq j-1$. We assign $q(\cdot|s_i,\cdot) := p^{s_0,y,\Delta}(\cdot|s_i,\cdot) \in \mathcal{P}_{s_i}$, which implies that $q_\Delta(s_{i+1}|s_i) \geq \delta/P_N$ for all $i \leq j-1$.

   Note that, at the current iteration, $q_\Delta(\cdot|s)$ is well-defined for all $s \in V$. Since $s_j \in V$, there is a non-repeating path $\{s_j = s'_j, s'_{j+1}, \ldots, s'_k = y\} \subset V$ s.t. $q_\Delta(s'_{i+1}|s'_i) \geq \delta/P_N$.

   Therefore, after assigning $q(\cdot|s_i,\cdot)$ for $i \leq j-1$, we have a new path $s_0 \to \ldots \to s_j \to s'_{j+1} \to \ldots \to s'_k = y$ with positive one step transition probabilities at least δ/P_N under q_Δ. We record this path that leads to y.

3. Update $V \leftarrow V \cup \{s_0, \ldots, s_{j-1}\}$.

Iterate until $V = S$.

Note that the algorithm terminates in at most $|S|$ iterations, producing $q \in \mathcal{P}$. Moreover, it produces a directed graph whose edges correspond to a positive transition probabilities at least δ/P_N under $q_\Delta(\cdot|\cdot)$, ensuring that every state can reach y in at most $|S|$ steps.

Therefore, we conclude that for any $y \in S$ and $\Delta : S \to \mathcal{Q}$, let $q = q^{y,\Delta} \in \mathcal{P}$ be constructed by the above algorithm, then

$$\min_{x\in S} P_x^{\Delta,q}(\tau_y \leq |S|) \geq (\delta/P_N)^{|S|} =: \delta' > 0.$$

Note that δ' is independent of x, y, Δ. This shows Lemma 2.

Step 3: We turn the probability bound in Lemma 2 to a bound on the expected hitting time bound.

Lemma 3. *Under the assumptions of Theorem 4, for any $\Delta : S \to \mathcal{Q}$ and $y \in S$, the $q \in \mathcal{P}$ in Lemma 2 satisfies*

$$\max_{x\in S} E_x^{\Delta,q}\tau_y \leq \frac{|S|}{\delta'}.$$

To show Lemma 3, we first show that

$$\max_{x\in S} P_x^{\Delta,q}(\tau_y > m|S|) \le (1-\delta')^m. \tag{B.3}$$

We prove by an induction on m. The base case $m=1$ follows directly from Lemma 2 that

$$\max_{x\in S} P_x^{\Delta,q}(\tau_y > |S|) = 1 - \min_{x\in S} P_x^{\Delta,q}(\tau_y \le |S|) \le 1-\delta'. \tag{B.4}$$

For the induction step, note that for any x

$$\begin{aligned}
P_x^{\Delta,q}(\tau_y > (k+1)|S|) &= E_x^{\Delta,q}\mathbb{1}\{\tau_y > (k+1)|S|\} \\
&= E_x^{\Delta,q}E_x^{\Delta,q}\left[\mathbb{1}\{\tau_y > (k+1)|S|\}\middle|\mathcal{H}_{k|S|}\right] \\
&\overset{(i)}{=} E_x^{\Delta,q}\left[\mathbb{1}\{\tau_y > k|S|\}\,E_x^{\Delta,q}\left[\mathbb{1}\{\tau_y > (k+1)|S|\}\middle|\mathcal{H}_{k|S|}\right]\right] \\
&\overset{(ii)}{=} E_x^{\Delta,q}\left[\mathbb{1}\{\tau_y > k|S|\}\,E_{X_{k|S|}}^{\Delta,q}\left[\mathbb{1}\{\tau_y > |S|\}\right]\right] \\
&= E_x^{\Delta,q}\left[\mathbb{1}\{\tau_y > k|S|\}\,P_{X_{k|S|}}^{\Delta,q}(\tau_y > |S|)\right] \\
&\overset{(iii)}{\le} E_x^{\Delta,q}[\mathbb{1}\{\tau_y > k|S|\}](1-\delta') \\
&\le (1-\delta')^{k+1}
\end{aligned}$$

where (i) follows from τ_y is a $\mathcal{H}_t$-stopping time with$\{\tau_y > k|S|\} = \{\tau_y \le k|S|\}^c \in \mathcal{H}_{k|S|}$, as well as $\mathbb{1}\{\tau_y > (k+1)|S|\} = \mathbb{1}\{\tau_y > (k+1)|S|\}\,\mathbb{1}\{\tau_y > k|S|\}$, (ii) is due to the Markov property, and (iii) follows from (B.4). This completes the induction step and shows (B.3).

We then prove Lemma 3 using (B.3). Note that since τ_y is non-negative, for $x \in S$, $s \ne y$,

$$\begin{aligned}
E_x^{\Delta,q}[\tau_y] &= \sum_{k\ge 0} P_x^{\Delta,q}(\tau_y \ge k) \\
&= \sum_{k\ge 0} P_x^{\Delta,q}(\tau_y > k) \\
&\le |S| + \sum_{k\ge 1} |S| P_x^{\Delta,q}(\tau_y \ge k|S|) \\
&\le |S| \sum_{k=0}^{\infty} (1-\delta')^k \\
&\le \frac{|S|}{\delta'}.
\end{aligned}$$

Of course $E_y^{\Delta,q}[\tau_y] = 0 \le |S|/\delta'$. This implies Lemma 3.

Step 4: Equipped Lemma 3, we now proceed to analyze the span of v_γ^*. We consider bounding $v_\gamma^*(x) - v_\gamma^*(y)$ for a fixed pair of states $x \ne y \in S$.

First, note that v_γ^* solves (3.3), then for each $\epsilon > 0$, there exists $\Delta_\epsilon : S \to \mathcal{Q}$ s.t. for all $s \in S$

$$v_\gamma^*(s) \le \inf_{p_s\in\mathcal{P}_s} E_{\Delta_\epsilon(\cdot|s),p_s}[r(s,A_0) + \gamma v_\gamma^*(X_1)] + (1-\gamma)\epsilon.$$

Then, by Theorem 1&5 in Wang et al. [7], for $\kappa \in \mathrm{K_H}$ denoting

$$v_\gamma^{\Delta_\epsilon,\kappa}(s) := E_s^{\Delta_\epsilon,\kappa} \sum_{k=0}^{\infty} \gamma^k r(X_k, A_k),$$

we have that for all $s \in S$,

$$0 \le v_\gamma^*(s) - \inf_{\kappa\in \mathrm{K_S}} v_\gamma^{\Delta_\epsilon,\kappa}(s) \le \epsilon;$$

i.e. Δ_ϵ is ϵ-optimal. Moreover, there exists stationary $p_\epsilon \in \mathrm{K_S}$ s.t.

$$0 \le v_\gamma^{\Delta_\epsilon, p_\epsilon}(s) - \inf_{\kappa \in \mathrm{K_S}} v_\gamma^{\Delta_\epsilon,\kappa}(s) \le \epsilon.$$

Note that by the definition of Δ_ϵ and p_ϵ,

$$\begin{aligned} v_\gamma^*(x) - v_\gamma^*(y) &= \sup_{\pi\in\Pi_\mathrm{S}} \inf_{\kappa\in\mathrm{K_H}} v_\gamma^{\pi,\kappa}(x) - \sup_{\pi\in\Pi_\mathrm{S}} \inf_{\kappa\in\mathrm{K_H}} v_\gamma^{\pi,\kappa}(y) \\ &\le \inf_{\kappa\in\mathrm{K_H}} v_\gamma^{\Delta_\epsilon,\kappa}(x) - \inf_{\kappa\in\mathrm{K_H}} v_\gamma^{\Delta_\epsilon,\kappa}(y) + \epsilon \\ &\le v_\gamma^{\Delta_\epsilon,\kappa}(x) - v_\gamma^{\Delta_\epsilon,p_\epsilon}(y) + 2\epsilon. \end{aligned} \tag{B.5}$$

for any $\kappa \in \mathrm{K_H}$. Since $\epsilon > 0$ can be arbitrarily small, it suffices to choose a $\kappa \in \mathrm{K_H}$ (potentially depending on $\Delta_\epsilon, p_\epsilon, x, y$) so that $v_\gamma^{\Delta_\epsilon,\kappa}(x) - v_\gamma^{\Delta_\epsilon,p_\epsilon}(y)$ is uniformly bounded.

This can be achieved using a similar argument as in Bartlett and Tewari [1]. We consider a history-dependent adversary $\kappa = (\kappa_0, \kappa_1, \dots) \in \mathrm{K_H}$ defined as follows. Let $g_{t-1} = (s_0, a_0, \dots, s_{t-1}, a_{t-1})$ and

$$\kappa_t(s'|g_{t-1}, s, a) = \begin{cases} q(s'|s,a) & \text{if } s_k \neq y, \forall k \le t-1 \text{ and } s \neq y, \\ p_\epsilon(s'|s,a) & \text{otherwise.} \end{cases} \tag{B.6}$$

Here $q = q^{y,\Delta_\epsilon}$ is defined in Lemma 2 and 3. In other words, the κ uses q when the chain hasn't hit y and uses the ϵ-optimal adversary after hitting y.

Under this history-dependent adversarial policy, we have

$$\begin{aligned} v_\gamma^{\Delta_\epsilon,\kappa}(x) &= E_x^{\Delta_\epsilon,\kappa} \sum_{k=0}^\infty \gamma^k r(X_k, A_k) \\ &= E_x^{\Delta_\epsilon,\kappa} \sum_{k=0}^{\tau_y - 1} \gamma^k r(X_k, A_k) + E_x^{\Delta_\epsilon,\kappa} \gamma^{\tau_y} \sum_{k=\tau_y}^\infty \gamma^{k-\tau_y} r(X_k, A_k) \\ &\le E_x^{\Delta_\epsilon,\kappa} \tau_y + E_x^{\Delta_\epsilon,\kappa} \sum_{k=\tau_y}^\infty \gamma^{k-\tau_y} r(X_k, A_k). \end{aligned} \tag{B.7}$$

Note that by the construction of κ in (B.6)

$$\begin{aligned} E_x^{\Delta_\epsilon,\kappa}\tau_y &= \sum_{k=0}^\infty P_x^{\Delta_\epsilon,\kappa}(\tau_y \ge k) \\ &\overset{(i)}{=} \sum_{k=0}^\infty P_x^{\Delta_\epsilon,\kappa}(\tau_y > k) \\ &= \sum_{k=0}^\infty E_x^{\Delta_\epsilon,\kappa} \mathbb{1}\{X_0(\omega), X_1(\omega), \dots, X_k(\omega) \neq y\} \\ &\overset{(ii)}{=} \sum_{k=0}^\infty \sum_{g_k = (s_0, a_0, \dots, s_k, a_k) \in \mathbf{G}_k} \prod_{j=0}^{k-1} \Delta_\epsilon(a_j|s_j)\kappa_j(s_{j+1}|g_{j-1}, s_j, a_j) \mathbb{1}\{s_0, \dots, s_k \neq y\} \\ &\overset{(iii)}{=} \sum_{k=0}^\infty \sum_{g_k \in \mathbf{G}_k} \prod_{j=0}^{k-1} \Delta_\epsilon(a_j|s_j) q(s_{j+1}|s_j, a_j) \mathbb{1}\{s_0, \dots, s_k \neq y\} \\ &= E_x^{\Delta_\epsilon, q}\tau_y. \end{aligned}$$

Here, (i) is becasue $x \neq y$, (ii) follows from the definition of $E_x^{\Delta_\epsilon,\kappa}$, (iii) is because by (B.6)

$$\kappa_j(s_{j+1}|g_{j-1}, s_j, a_j)\mathbb{1}\{s_0, \dots, s_k \neq y\} = \begin{cases} 0 & \text{if } s_i = y \text{ for some } i \le j \\ q(s_{j+1}|s_j, a_j) & \text{if } s_0, \dots, s_j \neq y \end{cases},$$

and the last equality follows from reversing the previous steps.

On the other hand,

$$
\begin{aligned}
&E_x^{\Delta_\epsilon,\kappa}\sum_{k=\tau_y}^{\infty}\gamma^{k-\tau_y}r(X_k,A_k)\\
&=E_x^{\Delta_\epsilon,\kappa}\sum_{j=1}^{\infty}\sum_{k=j}^{\infty}\mathbb{1}\left\{\tau_y=j\right\}\gamma^{k-\tau_y}r(X_k,A_k)\\
&\overset{(i)}{=}\sum_{j=1}^{\infty}\sum_{k=j}^{\infty}E_x^{\Delta_\epsilon,\kappa}\mathbb{1}\left\{\tau_y=j\right\}\gamma^{k-j}r(X_k,A_k)\\
&=\sum_{j=1}^{\infty}\sum_{k=j}^{\infty}\sum_{g_k\in\mathbf{G}_k}\Delta_\epsilon(a_j|s_j)\kappa_j(s_{j+1}|g_{j-1},s_j,a_j)\mathbb{1}\left\{s_0,\dots,s_{j-1}\neq y,s_j=y\right\}\gamma^{k-j}r(s_k,a_k)\\
&\overset{(ii)}{=}\sum_{j=1}^{\infty}\sum_{k=j}^{\infty}\sum_{g_k\in\mathbf{G}_k}\Delta_\epsilon(a_j|s_j)p_\epsilon(s_{j+1}|,s_j,a_j)\mathbb{1}\left\{s_0,\dots,s_{j-1}\neq y,s_j=y\right\}\gamma^{k-j}r(s_k,a_k)\\
&\overset{(iii)}{=}E_x^{\Delta_\epsilon,p_\epsilon}\sum_{k=\tau_y}^{\infty}\gamma^{k-\tau_y}r(X_k,A_k)\\
&=E_x^{\Delta_\epsilon,p_\epsilon}E_x^{\Delta_\epsilon,p_\epsilon}\left[\sum_{k=\tau_y}^{\infty}\gamma^{k-\tau_y}r(X_k,A_k)\middle|\mathcal{H}_{\tau_y}\right]\\
&\overset{(iv)}{=}E_y^{\Delta_\epsilon,p_\epsilon}\sum_{k=0}^{\infty}\gamma^{k}r(X_k,A_k)\\
&=v_\gamma^{\Delta_\epsilon,p_\epsilon}(y).
\end{aligned}
\tag{B.8}
$$

Here (i) applies Fubini's theorem leveraging the positivity of the summand, (ii) follows from the definition of κ in (B.6), (iii) is obtained from reversing the previous equalities, and (iv) is because p_ϵ is a stationary policy and, under $E_x^{\Delta_\epsilon,p_\epsilon}$, (X_t,A_t) is strong Markov.

Therefore, the bound in (B.5) and (B.7) implies that

$$
\begin{aligned}
v_\gamma^*(x)-v_\gamma^*(y)&\le v_\gamma^{\Delta_\epsilon,\kappa}(x)-v_\gamma^{\Delta_\epsilon,p_\epsilon}(y)+2\epsilon\\
&\le E_x^{\Delta_\epsilon,q}\tau_y+v_\gamma^{\Delta_\epsilon,p_\epsilon}(y)-v_\gamma^{\Delta_\epsilon,p_\epsilon}(y)+2\epsilon\\
&\le\frac{|S|}{\delta'}+2\epsilon.
\end{aligned}
$$

where the last inequality follows from Lemma 3.

Since $\epsilon>0$, $x,y\in S$ are arbitrary, we have that

$$
\left|v_\gamma^*\right|_{\text{span}}=\max_{x\in S}v_\gamma^*(x)-\min_{x\in S}v_\gamma^*(x)\le\frac{|S|}{\delta'}.
$$

So, $\left|v_\gamma^*\right|_{\text{span}}$ is uniformly bounded in γ. By Theorem 3, we conclude that (3.1) has a solution. This completes the proof of Theorem 4. □

B.3 Proof of Theorem 5

Proof. As in the proof of Theorem 4, we show that $\left|v_\gamma^*\right|_{\text{span}}$ is uniformly bounded in γ. To achieve this, we will be considering the inf-sup version of the γ-discounted Bellman equation defined as follows:

Definition 5. We say that $v'_\gamma:S\to\mathbb{R}$ is the unique solution to the γ-discounted inf-sup equation if

$$
v'_\gamma(s)=\inf_{p_s\in\mathcal{P}_s}\sup_{\phi\in\mathcal{Q}}E_{\phi,p_s}[r(s,A_0)+\gamma v'_\gamma(X_1)],\quad\forall s\in S.\tag{B.9}
$$

Similar to (3.3), the existence and uniqueness of the solutions of (B.9) follow from a standard contraction argument as in Wang et al. [7].

We note that under the convexity assumptions in Theorem 5, if v'_γ is the solution to (B.9), then an application of Sion's min-max theorem (see Corollary 2.1 in Wang et al. [7]) yields $v'_\gamma = v^*_\gamma$. Therefore, it suffices to show that $\left|v'_\gamma\right|_{\text{span}}$ is uniformly bounded in γ.

The remainder of the proof follows closely the argument of Theorem 4. In particular, modified versions of the key lemmas (with the roles of the controller and adversary reversed) continue to hold. Therefore, we provide only an abbreviated argument here, emphasizing the differences.

Step 1: The following lemma, analogous to Lemma 1, holds:

Lemma 4. *Under the assumptions of Theorem 5, there exists $\delta > 0$ s.t. for any stationary adversary policy $p \in \mathcal{P}$ and any pair of states $x, y \in S$, there exists Δ and $N \leq |S|$ s.t. $p^N_\Delta(y|x) \geq \delta$.*

We apply the same proof idea as before. For fixed $p \in \mathcal{P}$, by assumption, there exists $\Delta = \Delta^{x,y,p} \in \{S \to \mathcal{Q}\}$ so that $p^N_\Delta(y|x) > 0$ for some $N = N^{x,y,p}$. Observe that for Δ, N fixed, the mapping $p \to p^N_\Delta(y|x)$ is continuous. So, there exists an open neighborhood G_p of p s.t. $q^N_\Delta(y|x) \geq p^N_\Delta(y|x)/2$ for all $q \in G_p$.

Since $\mathcal{P}$ is compact, by the same argument as in the proof of Lemma 1, there exists a finite set $C := \left\{p_1, \ldots, p_{|C|}\right\} \subset \mathcal{P}$ s.t. $\{G_p : p \in C\}$ covers $\mathcal{P}$.

So, for any $q \in \mathcal{P}$, we can find $p \in C$ s.t. $q \in G_p$. Then, for any $x, y \in S$, set $\Delta = \Delta^{x,y,p}$ and $N = N^{x,y,p}$, we have

$$q^N_\Delta(y|x) \geq \frac{1}{2} p^N_\Delta(y|x) \geq \min_{p \in C} p^{N^{x,y,p}}_{\Delta^{x,y,p}}(y|x) =: \delta(y|x) > 0$$

where $\delta(y|x)$ is independent of p. Let $\delta = \min_{x,y \in S} \delta(y|x)$, this implies Lemma 4.

Step 2: We show the following lemma.

Lemma 5. *Under the assumptions of Theorem 5, there exists $\delta' > 0$ s.t. for any stationary adversary $p \in \mathcal{P}$ and states $y \in S$, there exists $\Delta : S \to \mathcal{Q}$ s.t.*

$$\min_{x \in S} P^{\Delta,p}_x(\tau_y \leq |S|) \geq \delta'.$$

For fixed adversary stationary policy $p \in \mathcal{P}$, we construct $\Delta = \Delta^{y,p}$ independent of x, that satisfies Lemma 5.

We will iteratively define $\Delta(\cdot|s) \in \mathcal{Q}$ until all $\{\Delta(\cdot|s) : s \in S\}$ has been assigned. Denote $p_\Delta(\cdot|s) = \sum_{a \in A} \Delta(a|s_i) p(\cdot|s, a)$.

We initialize the algorithm by assigning $\Delta(\cdot|y) = \phi(\cdot)$ for an arbitrary $\phi \in \mathcal{Q}$. Then, Let $V = \{y\}$ be the assigned states, and V^c the compliment in S are the unassigned states.

1. Choose any unassigned state $s_0 \in V^c$. Then by Lemma 1 there exists $\xi = \xi^{x,y,p} : S \to \mathcal{Q}$ and N s.t. $p^N_\xi(y|s_0) \geq \delta$. Therefore, by the "trimming" argument in the proof of Lemma 2, there exists a non-repeating path $s_0 \to s_1 \to, \ldots, \to s_k = y$, $s_i \neq s_j$ if $i \neq j$, s.t. $p_\xi(s_{i+1}|s_i) \geq \delta/P_N$ for all $i \leq k-1$, where $P_N := (|S|-1)!/(|S|-N)!$.
2. Let $j = \min\{k \geq 1 : s_k \in V\}$ be the first index so that s_j is assigned. So, $s_i \in V^c$ for all $i \leq j-1$. We assign $\Delta(\cdot|s_i) := \xi^{s_0,y,p}(\cdot|s_i) \in \mathcal{Q}$.

   Note that, at the current iteration, $\Delta(\cdot|s)$ is well-defined for all $s \in V$. Since $s_j \in V$, there is a non-repeating path $\left\{s_j = s'_j, s'_{j+1}, \ldots, s'_k = y\right\} \subset V$ s.t. $p_\Delta(s'_{i+1}|s'_i) \geq \delta/P_N$.

   Therefore, after assigning $\Delta(\cdot|s_i)$ for $i \leq j-1$, we have a new path $s_0 \to \ldots \to s_j \to s'_{j+1} \to \ldots \to s'_k = y$ with positive one step transition probabilities at least δ/P_N under p_Δ. We record this path that leads to y.
3. Update $V \leftarrow V \cup \{s_0, \ldots, s_{j-1}\}$.

Iterate until $V = S$.

This algorithm terminates in at most $|S|$ iterations, producing $\Delta : S \to \mathcal{Q}$. Moreover, it produces a directed graph whose edges correspond to positive transition probabilities at least δ/P_N under $p_\Delta(\cdot|\cdot)$, ensuring that every state can reach y in at most $|S|$ steps.

Therefore, we conclude that for any $y \in S$ and $p \in \mathcal{P}$, let Δ be constructed by the above algorithm, then

$$\min_{x\in S} P_x^{\Delta,p}(\tau_y \le |S|) \ge (\delta/P_N)^{|S|} =: \delta' > 0.$$

Note that δ' is independent of x, y, Δ. This shows Lemma 5.

Step 3: The same proof as that of Lemma 3 implies the following Lemma.

Lemma 6. *Under the assumptions of Theorem 5, for any $p \in \mathcal{P}$ and $y \in S$, the Δ constructed in Lemma 5 satisfies*

$$\max_{x\in S} E_x^{\Delta,p}\tau_y \le \frac{|S|}{\delta'}.$$

Step 4: By convexity of $\mathcal{Q}$ and $\mathcal{P}_s$ and the compactness of $\mathcal{P}_s$, Corollary 2.1 in Wang et al. [7] implies the existence of ϵ-optimal stationary strategies $\Delta : S \to \mathcal{Q}$ and $p_\epsilon \in \mathcal{P}$ of both the controller and the adversary. In particular,

$$\inf_{\kappa\in \mathrm{K_S}} v_\gamma^{\Delta_\epsilon,\kappa}(s) + \epsilon \ge v_\gamma^*(s) = v_\gamma'(s) \ge \sup_{\pi\in\Pi_\mathrm{H}} v_\gamma^{\pi,p_\epsilon}(s) - \epsilon.$$

Moreover, observe that

$$v_\gamma^{\Delta_\epsilon,\kappa_\epsilon}(s) + \epsilon \ge \inf_{\kappa\in \mathrm{K_S}} v_\gamma^{\Delta_\epsilon,\kappa}(s) + \epsilon \ge v_\gamma^*(s) \ge \sup_{\pi\in\Pi_\mathrm{H}} v_\gamma^{\pi,p_\epsilon}(s) - \epsilon \ge v_\gamma^{\Delta_\epsilon,\kappa_\epsilon}(s) - \epsilon.$$

So,

$$|v_\gamma^*(s) - v_\gamma^{\Delta_\epsilon,\kappa_\epsilon}(s)| \le \epsilon. \tag{B.10}$$

Next, following the same proof, we construct a two-phase history dependent controller policy by first using Δ constructed in 5 and 6 until the first time hitting y and then use an ϵ-optimal stationary policy Δ_ϵ afterwards. Let us denote this history-dependent controller by π.

By convexity of $\mathcal{Q}$ and $\mathcal{P}_s$ and the compactness of $\mathcal{P}_s$, Corollary 2.1 of Wang et al. [7] implies that

$$\begin{aligned} v_\gamma'(x) - v_\gamma'(y) &= \inf_{\kappa\in \mathrm{K_S}} \sup_{\pi\in\Pi_\mathrm{H}} v_\gamma^{\pi,p}(x) - \inf_{\kappa\in \mathrm{K_S}} \sup_{\pi\in\Pi_\mathrm{H}} v_\gamma^{\pi,\kappa}(y) \\ &\ge \sup_{\pi\in\Pi_\mathrm{H}} v_\gamma^{\pi,p_\epsilon}(x) - \sup_{\pi\in\Pi_\mathrm{H}} v_\gamma^{\pi,p_\epsilon}(y) - \epsilon \\ &\ge v_\gamma^{\pi,p_\epsilon}(x) - v_\gamma^{\Delta_\epsilon,p_\epsilon}(x) - 3\epsilon, \end{aligned} \tag{B.11}$$

where the last inequality follows from (B.10).

We observe that

$$\begin{aligned} v_\gamma^{\pi,p_\epsilon}(x) &= E_x^{\pi,p_\epsilon} \sum_{k=0}^\infty \gamma^k r(X_k, A_k) \\ &= E_x^{\pi,p_\epsilon} \sum_{k=0}^{\tau_y-1} \gamma^k r(X_k, A_k) + E_x^{\pi,p_\epsilon} \gamma^{\tau_y} \sum_{k=\tau_y}^\infty \gamma^{k-\tau_y} r(X_k, A_k) \\ &\ge E_x^{\pi,p_\epsilon}(\gamma^{\tau_y} - 1) \sum_{k=\tau_y}^\infty \gamma^{k-\tau_y} r(X_k, A_k) + E_x^{\pi,p_\epsilon} \sum_{k=\tau_y}^\infty \gamma^{k-\tau_y} r(X_k, A_k). \end{aligned}$$

Note that

$$
\begin{aligned}
E_x^{\pi,p_\epsilon}(\gamma^{\tau_y}-1)\sum_{k=\tau_y}^{\infty}\gamma^{k-\tau_y}r(X_k,A_k) &\geq E_x^{\pi,p_\epsilon}\frac{\gamma^{\tau_y}-1}{1-\gamma}\\
&= -E_x^{\pi,p_\epsilon}\sum_{k=0}^{\tau_y-1}\gamma^k\\
&\geq -E_x^{\pi,p_\epsilon}\tau_y\\
&= -E_x^{\Delta,p_\epsilon}\tau_y\\
&\geq -\frac{|S|}{\delta'}
\end{aligned}
$$

and by (B.8),

$$
E_x^{\pi,p_\epsilon}\sum_{k=\tau_y}^{\infty}\gamma^{k-\tau_y}r(X_k,A_k) = v_\gamma^{\Delta_\epsilon,p_\epsilon}(y).
$$

Therefore, from (B.11), we have that

$$
v'_\gamma(x) - v'_\gamma(y) \geq -\frac{|S|}{\delta'} - 3\epsilon.
$$

This hold for all $x, y \in S$, and ϵ can be arbitrarily small. Hence, we have that $|v'_\gamma(x)-v'_\gamma(y)| \leq |S|/\delta'$ for all $x, y \in S$.

Under the assumptions of Theorem 5, Corollary 2.1 in Wang et al. [7] implies that $v'_\gamma = v^*_\gamma$ solves (3.3) and (B.9). So, we conclude that

$$
\left|v'_\gamma\right|_{\text{span}} = \left|v^*_\gamma\right|_{\text{span}} \leq \frac{|S|}{\delta'}
$$

which is uniform in γ. This and Theorem 3 implies Theorem 5. □

C Proofs for Section 5

C.1 Proof of Proposition 5.1

Proof. First, notice that

$$
\begin{aligned}
0 &\leq \underline{\alpha}(\mu,\Pi_{\text{H}},\text{K}_{\text{S}}) - \inf_{\kappa\in\text{K}_{\text{S}}}\underline{\alpha}(\mu,\pi_{\text{RL}},\kappa)\\
&\overset{(i)}{\leq} \inf_{\kappa\in\text{K}_{\text{S}}}\sup_{\pi\in\Pi_{\text{H}}}\underline{\alpha}(\mu,\pi,\kappa) - \inf_{\kappa\in\text{K}_{\text{S}}}\underline{\alpha}(\mu,\pi_{\text{RL}},\kappa)\\
&\overset{(ii)}{=} \inf_{\kappa\in\text{K}_{\text{S}}}\sup_{\pi\in\Pi_{\text{S}}}\underline{\alpha}(\mu,\pi,\kappa) - \inf_{\kappa\in\text{K}_{\text{S}}}\underline{\alpha}(\mu,\pi_{\text{RL}},\kappa)\\
&\leq \inf_{\kappa\in\text{K}_{\text{S}}}\left|\sup_{\pi\in\Pi_{\text{S}}}\underline{\alpha}(\mu,\pi,\kappa) - \underline{\alpha}(\mu,\pi_{\text{RL}},\kappa)\right|\\
&\overset{(iii)}{=} \inf_{\kappa\in\text{K}_{\text{S}}}|\alpha^*_\kappa - \underline{\alpha}(\mu,\pi_{\text{RL}},\kappa)|
\end{aligned}
\tag{C.1}
$$

where (i) follows from weak duality and (ii) uses the optimality of Π_{S} for classical MDPs (see Puterman [6]). For (iii), note that since $p \in \mathcal{P}$ is weakly communicating, by the standard results from classical MDPs (also see Puterman [6]), we have that for each $\kappa \in \text{K}_{\text{S}}$, there exists an optimal deterministic Markov time-homogeneous policy Δ_κ that achieves an optimal average-reward α^*_κ.

On the other hand, Algorithm 2 in Zhang and Xie [11] and the regret bound therein imply that for any weakly communicating MDP and parameter $\epsilon > 0$, there exists a policy π_{RL} that uses only deterministic actions so that for any $\kappa \in \text{K}_{\text{S}}$, w.p. at least $1-\epsilon$

$$
\sum_{t=0}^{n-1}[\alpha^*_\kappa - r(X_t,A_t)] = \widetilde{O}(|h^*_\kappa|_{\text{span}}\sqrt{n}).
$$

This implies that

$$0 \le \alpha_\kappa^* - E_\mu^{\pi,\kappa} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t) = \widetilde{O}\left(\frac{|h_\kappa^*|_{\text{span}}}{\sqrt{n}}\right)(1-\epsilon) + \epsilon.$$

Hence, we have that

$$\begin{aligned} 0 &\le \alpha_\kappa^* - \overline{\alpha}(\mu, \pi_{\text{RL}}, \kappa) \\ &\le \alpha_\kappa^* - \underline{\alpha}(\mu, \pi_{\text{RL}}, \kappa) \\ &= \limsup_{n\to\infty} \left(\alpha_\kappa^* - E_\mu^{\pi,\kappa} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t)\right) \\ &\le \epsilon \end{aligned} \tag{C.2}$$

Since π_{RL} only uses deterministic actions and $\{\delta_a : a \in A\} \subset \mathcal{Q}$, $\pi_{\text{RL}} \in \Pi_{\text{H}}(\mathcal{Q})$. Therefore, going back to (C.1), we have that

$$\begin{aligned} 0 &\le \underline{\alpha}(\mu, \Pi_{\text{H}}, \text{K}_{\text{S}}) - \inf_{\kappa \in \text{K}_{\text{S}}} \underline{\alpha}(\mu, \pi_{\text{RL}}, \kappa) \\ &\le \inf_{\kappa \in \text{K}_{\text{S}}} \sup_{\pi \in \Pi_{\text{S}}} \underline{\alpha}(\mu, \pi, \kappa) - \inf_{\kappa \in \text{K}_{\text{S}}} \underline{\alpha}(\mu, \pi_{\text{RL}}, \kappa) \\ &\le \inf_{\kappa \in \text{K}_{\text{S}}} |\alpha_\kappa^* - \underline{\alpha}(\mu, \pi_{\text{RL}}, \kappa)| \\ &\le \epsilon. \end{aligned}$$

Finally, to conclude the proposition, we note that if $\underline{\alpha}$ is replaced by $\overline{\alpha}$, the derivation in (C.1) is still valid. This, coupled with (C.2), yields the limsup version of Theorem 5.1. □

C.2 Proof of Theorem 6

Proof. By Proposition 5.1, we have that for any $\epsilon > 0$,

$$-\epsilon \le \underline{\alpha}(\mu, \Pi_{\text{H}}, \text{K}_{\text{S}}) - \inf_{\kappa \in \text{K}_{\text{S}}} \sup_{\pi \in \Pi_{\text{S}}} \underline{\alpha}(\mu, \pi, \kappa) \le \epsilon.$$

Also, by Markov optimality in classical MDPs [6], $\sup_{\pi \in \Pi_{\text{S}}} \underline{\alpha}(\mu, \pi, \kappa) = \sup_{\pi \in \Pi_{\text{H}}} \underline{\alpha}(\mu, \pi, \kappa)$. Since ϵ can be arbitrarily small, these inequalities imply the liminf version of Theorem 6. The same argument holds when $\underline{\alpha}$ is replaced with $\overline{\alpha}$.

To show the second claim, we note that the same argument as in the proof of Theorem 3 implies that if $\left|v_\gamma'\right|_{\text{span}}$ is uniformly bounded in γ, then (3.2) has a solution (u', α'). Moreover, the argument for Theorem 1 will imply that α' is the optimal average-reward

$$\alpha' = \inf_{\kappa \in \text{K}_{\text{H}}} \sup_{\pi \in \Pi_{\text{H}}} \underline{\alpha}(\mu, \pi, \kappa) = \inf_{\kappa \in \text{K}_{\text{S}}} \sup_{\pi \in \Pi_{\text{H}}} \underline{\alpha}(\mu, \pi, \kappa) = \inf_{\kappa \in \text{K}_{\text{S}}} \sup_{\pi \in \Pi_{\text{S}}} \underline{\alpha}(\mu, \pi, \kappa).$$

So, we only need to show that $\left|v_\gamma'\right|_{\text{span}}$ is uniformly bounded in γ. This is achieved by employing the same proof as Theorem 4. As the proof idea is already used in both Theorem 4 and 5, we omit further detail. □

C.3 Proof of Corollary 6.1

Proof. By Theorem 4 and 6, solutions (u^*, α^*) and (u', α') to (3.1) and (3.2) exists under the assumptions of Corollary 6.1. Hence, by Theorem 1 and 6,

$$\underline{\alpha}(\mu, \Pi_{\text{H}}, \text{K}_{\text{S}}) \;=\; \overline{\alpha}(\mu, \Pi_{\text{H}}, \text{K}_{\text{S}}) \;=\; \alpha',$$

while

$$\underline{\alpha}(\mu, \Pi_{\text{S}}, \text{K}_{\text{S}}) \;=\; \overline{\alpha}(\mu, \Pi_{\text{S}}, \text{K}_{\text{S}}) \;=\; \alpha^*.$$

This implies the corollary. □