DISTRIBUTIONALLY ROBUST AVERAGE-REWARD REINFORCEMENT LEARNING: FINITE-SAMPLE GUARANTEES UNDER WEAK COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study distributionally robust reinforcement learning (DR-RL) in the average-reward setting under weak communication. Our main result provides finite-sample guarantees for estimating the robust optimal average reward and learning a near-optimal policy, covering both SA-rectangular and S-rectangular structures with divergence-based and distance-based uncertainty sets. Specifically, for Kullback–Leibler and f_k-divergence balls, we establish explicit radius conditions under which the robust average-reward Bellman equation admits a constant-gain solution, while for total variation and Wasserstein balls, any positive radius suffices without requiring the nominal MDP is weakly communicating. Our algorithm is prior-knowledge-free and achieves sample complexities of $\widetilde{O}(|\mathcal{S}||\mathcal{A}|p_\wedge^{-1}\,\mathrm{Span}^2(u_\delta^*)\epsilon^{-2})$ for estimating the robust optimal average reward and $\widetilde{O}(|\mathcal{S}||\mathcal{A}|p_\wedge^{-2}\,\mathrm{Span}^2(u_\delta^*)\epsilon^{-2})$ for learning an ϵ-optimal policy. Here, $p_\wedge$ is the smallest positive nominal transition probability and u_δ^* is a robust optimal bias function. We further provide an almost-tight explicit upper bound on $\mathrm{Span}(u_\delta^*)$. Finally, we validate the predicted $n^{-1/2}$ convergence rate through numerical experiments.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 2018) is a core machine learning framework in which agents learn to maximize long-term rewards by interacting with their environments. RL has achieved substantial success across a wide range of applications, from robotics and control (Kober et al., 2013; Gu et al., 2017) to recent advances in reasoning with large language models (LLMs) (Wei et al., 2022; Guo et al., 2025).

Despite these successes, RL faces a fundamental challenge when the transition dynamics encountered during deployment differ from those observed during training. Such discrepancies can arise from changes in operating conditions, unmodeled disturbances, or adversarial perturbations, potentially degrading the performance of policies learned under a nominal model.

Distributionally robust reinforcement learning (DR-RL) addresses this challenge by explicitly accounting for uncertainty in transition dynamics. Built on the robust Markov decision process framework (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013), DR-RL seeks policies that perform well under the worst-case transition model within a prescribed uncertainty set. Subsequent work has developed learning algorithms and theoretical guarantees in both model-based (Panaganti & Kalathil, 2022; Xu et al., 2023; Clavier et al., 2024b; Shi et al., 2026) and model-free (Liu et al., 2022; Wang et al., 2023a; 2024b) settings.

Two widely studied uncertainty structures are SA-rectangularity and S-rectangularity, which specify how the adversary's choices are coupled across states and actions (Le Tallec, 2007; Wiesemann et al., 2013). Under SA-rectangularity, the adversary selects transition distributions independently for each state-action pair, and SA-rectangular models always admit a deterministic and stationary optimal policy. Under S-rectangularity, these choices remain independent across states but may be coupled across actions within each state, the optimal policy may be history-dependent and randomized (Grand-Clément et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Much of the existing finite-sample distributionally robust reinforcement learning theory concerns discounted problems (Li et al., 2026; Shi et al., 2026; Clavier et al., 2024b;a; Panaganti & Kalathil, 2022; Wang et al., 2024b; 2023a), in which future rewards are geometrically discounted. For continuing tasks evaluated by long-run performance, the average-reward criterion provides a natural alternative. Existing sample complexity guarantees for average-reward DR-RL primarily concern SA-rectangular uncertainty under strong ergodic assumptions (Chen et al., 2025; Roch et al., 2026; Xu et al., 2026; 2025). Despite progress on structural results under weaker assumptions (Wang & Si, 2025), finite-sample guarantees under weak communication assumptions remain, to the best of our knowledge, unavailable for both SA- and S-rectangular uncertainty.

In this work, our main contribution are summarized as follows.

- **Structural guarantees under weak communication.** We establish sufficient conditions for the existence of a constant-gain robust average-reward Bellman solution under both SA- and S-rectangular uncertainty. For KL- and f_k-divergence balls, nominal weak communication and suitable radius restrictions (Assumption 1) preserve the nominal transition support, ensuring that every feasible adversarial kernel remains weakly communicating (Proposition 1). For TV and Wasserstein balls, any positive radius suffices to guarantee a constant-gain Bellman solution without requiring nominal weak communication (Proposition 2). Under the same conditions, a robust average-reward optimal policy can be chosen stationary: deterministic under SA-rectangularity and randomized under S-rectangularity (Corollary 1).
- **Unified finite-sample guarantees for gain estimation and policy learning.** We develop a prior knowledge free algorithm via discounted reduction (Algorithm 1) covering all eight combinations of rectangularity structure and uncertainty families (KL, f_k, TV, and Wasserstein). Our analysis establishes Bellman-operator perturbation bounds that hold uniformly over stationary randomized policies and value functions, and controls the discounted-to-average-reward approximation through the robust optimal bias span. We obtain an ε-accurate estimate of the robust optimal average reward and optimal policy using
$$\widetilde{O}(|\mathcal{S}||\mathcal{A}|p_\wedge^{-1}\,\mathrm{Span}(u_\delta^*)^2\varepsilon^{-2}), \quad \widetilde{O}(|\mathcal{S}||\mathcal{A}|p_\wedge^{-2}\,\mathrm{Span}(u_\delta^*)^2\varepsilon^{-2})$$
samples respectively (Theorem 1), where $p_\wedge$ is the smallest positive nominal transition probability and u_δ^* is a robust optimal bias function, and $\mathrm{Span}(u_\delta^*) := \max_{s\in\mathcal{S}} u_\delta^*(s) - \min_{s\in\mathcal{S}} u_\delta^*(s)$.
- **Explicit bias-span bounds and sharpness.** We derive explicit bounds on $\mathrm{Span}(u_\delta^*)$ for KL, f_k, TV, and Wasserstein uncertainty under both rectangularity structures, making our span-dependent policy guarantee explicit (Theorem 2). The TV and Wasserstein bounds are attained by two-state instances; for KL and f_k, our construction shows that the exponent of $p_\wedge^{-1}$ cannot be improved (Theorem 3).

The remainder of this paper is organized as follows. Section 2 reviews related work, with an emphasis on average-reward DR-RL. Section 3 introduces the notation and problem formulation. Section 4 presents our main theoretical results, including the algorithm and its sample complexity analysis. Section 5 presents numerical experiments illustrating our theoretical findings.

2 Literature Review

Average-reward RL. Average-reward RL provides a classical framework for sequential decision-making under long-run performance criteria (Puterman, 2005). In the non-robust setting, model-based approaches have been extensively studied from the perspectives of regret minimization and exploration (Jaksch et al., 2010; Fruit et al., 2018), as well as through structural complexity measures such as mixing time, minorization time, and bias span (Jin & Sidford, 2021; Wang et al., 2024a; Zurek & Chen, 2024; 2025b;a). Beyond model-based approaches, a growing body of work has investigated model-free methods, including Q-learning and its variants (Jin & Sidford, 2020; Zhang et al., 2021; Chen, 2025; Chen et al., 2026; Zhang & Xie, 2023; Agrawal & Agrawal, 2025; Lee et al., 2025; 2026).

Distributionally Robust Average-Reward RL: Grand-Clément et al. (2026) establish conditions under which a stationary optimal policy exists. On the algorithmic side, Wang et al. (2023c;b; 2024c)

develop robust relative value iteration and TD/Q-learning algorithms and establish their convergence in the distributionally robust average-reward setting. Roch & Wang (2026) study average-reward distributionally robust Markov games, establishing Nash equilibrium existence under irreducibility and weak communication conditions and developing algorithms with convergence guarantees. Moving beyond convergence analysis, subsequent work provides sample complexity guarantees under SA-rectangular uncertainty, with assumptions ranging from irreducibility and aperiodicity to the unichain condition (Xu et al., 2025; 2026; Roch et al., 2026; Chen et al., 2025; Yang et al., 2026). For s-rectangular uncertainty, Wang & Si (2025) extend the structural theory by providing weak communication conditions ensuring the existence of solutions to the constant-gain robust Bellman equation. Despite these advances, a gap remains between structural results and finite-sample analysis under weak communication assumptions: to the best of our knowledge, non-asymptotic sample complexity guarantees in this setting remain open for both SA and S-rectangular uncertainty.

We provide a summary of state-of-the-art sample complexity results in the literature in Table 1.

Table 1: Summary of DR-AMDP sample complexity bounds under a generative model. Here, $\kappa \in (0,1)$ denotes the robust Bellman contraction factor under the specific seminorm defined in Xu et al. (2025), $\mathcal{H}$ is the maximum of bias-span over all optimal policies and corresponding worst-case kernel which is at least $\mathrm{Span}(u_\delta^*)$. t_{mix} denotes the mixing-time bound used in the corresponding cited result.

Setting	Uncertainty Type	Sample Complexity	Assumption	Origin
SA-rectangular	TV, Wasserstein	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert t_{\mathrm{mix}}^2(1-\kappa)^{-2}\varepsilon^{-2}\right)$	Irreducible and Aperiodic	Xu et al. (2025)
	TV, Wasserstein	$\widetilde{O}\left(\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2\varepsilon^{-2}\right)$	Irreducible and Aperiodic	Xu et al. (2026)
	KL, χ^2	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\mathcal{H}^2\varepsilon^{-(2+o(1))}\right)$	Irreducible	Roch et al. (2026)
	KL, f_k	$\widetilde{O}(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert t_{\mathrm{mix}}^2 p_\wedge^{-1}\varepsilon^{-2})$	Uniformly ergodic	Chen et al. (2025)
	TV	$\widetilde{O}(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert(\mathrm{Span}(u_\delta^*)+\delta\,\mathrm{Span}(u_\delta^*)^2)\varepsilon^{-2})$	Unichain	Yang et al. (2026)
	KL, f_k, TV, Wasserstein	$\widetilde{O}(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert p_\wedge^{-2}\,\mathrm{Span}(u_\delta^*)^2\varepsilon^{-2})$	Weakly communicating	**This work**
S-rectangular	KL, f_k, TV, Wasserstein	$\widetilde{O}\left(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert p_\wedge^{-2}\,\mathrm{Span}(u_\delta^*)^2\varepsilon^{-2}\right)$	Weakly communicating	**This work**

3 Preliminaries and Problem Formulation

3.1 Markov Decision Process

We briefly review the classical finite tabular MDP models and introduce the notation used throughout the paper. Let $\Delta(\mathcal{S})$, $\Delta(\mathcal{A})$ denote the probability simplex over the finite state space $\mathcal{S}$ and action space $\mathcal{A}$ respectively. $P := \{p_{s,a} \in \Delta(\mathcal{S}) : (s,a) \in \mathcal{S}\times\mathcal{A}\}$ is the controlled transition kernel, where $p_{s,a}(s')$ denotes the probability of transitioning to state s' after action a is selected in state s. $r : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$ is the reward function. A finite discounted MDP (DMDP) is specified by the tuple $\mathcal{M}_\gamma = (\mathcal{S},\mathcal{A},P,r,\gamma)$ where $\gamma \in (0,1)$ is the discount factor, and an average-reward MDP (AMDP) is specified as $\mathcal{M} = (\mathcal{S},\mathcal{A},P,r)$.

Let Π_{HD} denote the class of history-dependent randomized policies $\pi = (\pi_t)_{t\geq 0}$, where $\pi_t(\cdot \mid h_t) \in \Delta(\mathcal{A})$ and $h_t = (s_0,a_0,\ldots,s_t)$ is the observed history. Stationary randomized policies depend only on the current state and form $\Pi_{\mathrm{SR}} := \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$. Stationary deterministic policies select a single action in each state and form $\Pi_{\mathrm{SD}} := \{\pi : \mathcal{S} \to \mathcal{A}\}$. For ordinary finite DMDPs and AMDPs, it suffices to consider stationary deterministic policies (Puterman, 2005, Chapters 6 and 8). Identifying deterministic policies with point-mass action distributions gives $\Pi_{\mathrm{SD}} \subseteq \Pi_{\mathrm{SR}} \subseteq \Pi_{\mathrm{HD}}$. For $\pi \in \Pi_{\mathrm{SR}}$, the induced nominal transition kernel is $P_\pi(x \mid s) := \sum_{a\in\mathcal{A}} \pi(a \mid s) p_{s,a}(x)$.

Discounted criterion. For a given discounted MDP (DMDP) instance $\mathcal{M}_\gamma = (\mathcal{S},\mathcal{A},P,r,\gamma)$ and policy $\pi \in \Pi_{\mathrm{SD}}$, the *discounted value function* $V_\gamma^\pi(s) := \mathrm{E}_P^\pi\left[\sum_{t=0}^\infty \gamma^t r(S_t,A_t,S_{t+1}) \mid S_0 = s\right]$. An optimal policy $\pi^* \in \Pi_{\mathrm{SD}}$ achieves the optimal value $V_\gamma^*(s) := \max_{\pi\in\Pi_{\mathrm{SD}}} V_\gamma^\pi(s)$.

Average-reward criterion. An average-reward MDP (AMDP) instance $\mathcal{M} = (\mathcal{S},\mathcal{A},P,r)$ and policy $\pi \in \Pi_{\mathrm{SD}}$, the *long-term average-reward function* (gain) $g^\pi : \mathcal{S} \to \mathbb{R}$ is defined as $g^\pi(s) := \limsup_{T\to\infty} T^{-1}\mathrm{E}_P^\pi[\sum_{t=0}^{T-1} r(S_t,A_t,S_{t+1}) \mid S_0 = s]$ and optimal gain $g^*(s) := \sup_{\pi\in\Pi_{\mathrm{SD}}} g^\pi(s)$ is a constant function when P is weakly communicating (Definition 1).

For average-reward MDPs, the following weak communication condition ensures a state-independent optimal gain without requiring every stationary policy to induce an ergodic chain (Puterman, 2005).

Definition 1 (Weakly Communicating)**.** An MDP is weakly communicating if there exists a partition $\mathcal{S} = C \cup T$ such that every pair of states in C communicates under some stationary policy $\pi \in \Pi_{\mathrm{SR}}$, and every state in T is transient under every stationary policy $\pi \in \Pi_{\mathrm{SR}}$.

3.2 Robust MDPs and Rectangularity

We study robust MDPs under both SA- and S-rectangular uncertainty, where an adversary perturbs the nominal transition probabilities within a prescribed uncertainty set. We consider divergence-based sets defined by Kullback–Leibler (KL) and f_k divergences, and distance-based sets defined by total variation (TV) and Wasserstein distances. We use $U \in \{D_{\mathrm{KL}}, D_{f_k}, d_{\mathrm{TV}}, W_\ell\}$ to specify the uncertainty measure and δ for its radius. For $q, p \in \Delta(\mathcal{S})$ with $q \ll p$, define

$$D_{\mathrm{KL}}(q\|p) := \sum_{s \in \mathrm{supp}(p)} q(s) \log \frac{q(s)}{p(s)}, \qquad D_{f_k}(q\|p) := \sum_{s \in \mathrm{supp}(p)} p(s) f_k\left(\frac{q(s)}{p(s)}\right),$$

where $0 \log 0 := 0$ and $f_k(t) := (t^k - kt + k - 1)/(k(k-1))$ for $t \geq 0$ and $k \in (1, \infty)$. For arbitrary $q, p \in \Delta(\mathcal{S})$, define

$$d_{\mathrm{TV}}(q, p) := \frac{1}{2}\|q - p\|_1, \qquad W_\ell(q, p) := \left(\inf_{\xi \in \Gamma(q,p)} \sum_{x,y \in \mathcal{S}} \rho(x, y)^\ell \xi(x, y)\right)^{1/\ell},$$

where $\Gamma(q, p)$ denotes the set of couplings of q and p, ρ is a metric on $\mathcal{S}$, and $\ell \in [1, \infty)$. Unlike the divergence-based sets considered here, the distance-based sets do not require absolute continuity with respect to the nominal transition distribution. When U is a divergence, $U(q, p) = D(q\|p)$.

The statistical complexity of policy learning in robust MDPs are primarily studied under SA- and S-rectangular uncertainty sets. SA-rectangularity allows the adversary to choose each state-action pair's transition distribution in- dependently. Under S-rectangularity, the constraints separate only across states, while the transition distributions for different actions at the same state may be jointly constrained. From this point forward, we introduce SA- and S-rectangular respectively.

Definition 2 (SA-rectangularity)**.** The uncertainty set $\mathcal{P}$ is called an SA-rectangular set if it satisfies $\mathcal{P} = \prod_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{P}_{s,a}(U, \delta)$ where $\mathcal{P}_{s,a} \subseteq \Delta(\mathcal{S})$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Specifically, for divergence-based uncertainty set $U \in \{D_{\mathrm{KL}}, D_{f_k}\}$, we set

$$\mathcal{P}_{s,a}(U, \delta) := \{q_{s,a} \in \Delta(\mathcal{S}) : q_{s,a} \ll p_{s,a},\ U(q_{s,a}\|p_{s,a}) \leq \delta\}.$$

For distance-based uncertainty $U \in \{d_{\mathrm{TV}}, W_\ell\}$, $\ll$ is not required and

$$\mathcal{P}_{s,a}(U, \delta) := \{q_{s,a} \in \Delta(\mathcal{S}) : U(q_{s,a}, p_{s,a}) \leq \delta\}.$$

In addition, Wiesemann et al. (2013) propose the extensive version S-rectangular set which is detained in Definition 3.

Definition 3 (S-rectangularity)**.** The uncertainty set $\mathcal{P}$ is called an S-rectangular set if it satisfies $\mathcal{P} = \prod_{s \in \mathcal{S}} \mathcal{P}_s(U, \delta)$ where $\mathcal{P}_s \subseteq \Delta(\mathcal{S})^{|\mathcal{A}|}$ and $\Delta(\mathcal{S})^{|\mathcal{A}|} := \{q_s = (q_{s,a})_{a \in \mathcal{A}} | q_{s,a} \in \Delta(\mathcal{S}),\ \text{for all } a \in \mathcal{A}\}$.

where $\mathcal{P}_s(U, \delta)$ jointly constrains the collection $(q_{s,a})_{a \in \mathcal{A}}$ at state s. In this paper, we consider the sum-form of S-rectangularity, which is commonly studied in the literature; see, for example, Yang et al. (2022); Li et al. (2026). In particular, for divergence-based uncertainty, $U \in \{D_{\mathrm{KL}}, D_{f_k}\}$, we set

$$\mathcal{P}_s(U, \delta) := \left\{q_s = (q_{s,a})_{a \in \mathcal{A}} : q_{s,a} \ll p_{s,a}, \sum_{a \in \mathcal{A}} U(q_{s,a}\|p_{s,a}) \leq |\mathcal{A}|\delta\right\}$$

For distance-based uncertainty $\mathcal{P}_s(U, \delta)$ where $U \in \{d_{\mathrm{TV}}, W_\ell\}$, the budget constraint is

$$\sum_{a \in \mathcal{A}} d_{\mathrm{TV}}(q_{s,a}, p_{s,a}) \leq |\mathcal{A}|\delta, \qquad \sum_{a \in \mathcal{A}} W_\ell(q_{s,a}, p_{s,a})^\ell \leq |\mathcal{A}|\delta^\ell.$$

For SA-rectangularity, take $\mathcal{P}_s = \prod_{a\in\mathcal{A}} \mathcal{P}_{s,a}$ in the following definition.

Definition 4. The discounted and average-reward optimal Bellman operators $\mathcal{T}_\gamma^*, \mathcal{T}^* : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ are defined by

$$\mathcal{T}_\gamma^*(v)(s) := \max_{\phi\in\Delta(\mathcal{A})} \inf_{q_s\in\mathcal{P}_s} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \phi(a) q_{s,a}(s') \big(r(s,a,s') + \gamma v(s')\big), \tag{3.1}$$

$$\mathcal{T}^*(v)(s) := \max_{\phi\in\Delta(\mathcal{A})} \inf_{q_s\in\mathcal{P}_s} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \phi(a) q_{s,a}(s') \big(r(s,a,s') + v(s')\big). \tag{3.2}$$

Under SA-rectangularity, the adversary minimizes each action independently, so these operators reduce to $\mathcal{T}_\gamma^*(v)(s) = \max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\mathcal{P}_{s,a}} \sum_{s'\in\mathcal{S}} q_{s,a}(s') \big(r(s,a,s') + \gamma v(s')\big)$ and $\mathcal{T}^*(v)(s) = \max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\mathcal{P}_{s,a}} \sum_{s'\in\mathcal{S}} q_{s,a}(s') \big(r(s,a,s') + v(s')\big)$ (Iyengar, 2005; Nilim & El Ghaoui, 2005). Under S-rectangularity, randomization may be necessary (Wiesemann et al., 2013; Wang & Si, 2025).

It is well known that, under either SA- or S-rectangularity, the discounted Bellman equation associated with $\mathcal{T}_\gamma^*$ in (3.1) admits a unique fixed point v_γ^*, i.e., $v_\gamma^* = \mathcal{T}_\gamma^*(v_\gamma^*)$ which equals the robust optimal discounted value (Iyengar, 2005; Wiesemann et al., 2013).

3.3 Long-Run Structure: Gain and Bias Span

We now turn to long-run performance under transition uncertainty. In the average-reward setting, the controller uses a history-dependent policy, while the adversary may choose admissible transition kernels based on the observed history. For a fixed $\pi \in \Pi_{\mathrm{HD}}$, let $\mathbf{Q} = (Q_t)_{t\geq 0}$ denote an admissible sequence of transition kernels, where $Q_t \in \mathcal{P}$ may depend on the history up to time t. The robust average reward of π is defined by

$$g_\delta^\pi(s) := \inf_{\mathbf{Q}} \limsup_{T\to\infty} \mathrm{E}_{\mathbf{Q}}^\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} r(S_t, A_t, S_{t+1}) \,\middle|\, S_0 = s \right], \qquad s \in \mathcal{S}.$$

The robust optimal gain is then $g_\delta^*(s) := \sup_{\pi\in\Pi_{\mathrm{HD}}} g_\delta^\pi(s)$. When $g_\delta^*(s)$ is independent of the initial state, we denote the common value by g_δ^*.

Wang & Si (2025) establish the existence of a constant-gain robust Bellman solution under adversary-side weak communication of the uncertainty set, together with suitable compactness and minimax conditions. For divergence-based uncertainty sets, this requirement on the admissible kernels $Q \in \mathcal{P}$ does not directly translate into conditions on the nominal kernel P and radius δ. We therefore seek such conditions for the same constant-gain Bellman equation, written with our average-reward optimal Bellman operator $\mathcal{T}^*$ as

$$\mathcal{T}^*(u_\delta^*) = u_\delta^* + g\mathbf{1}. \tag{3.3}$$

We assume $\mathrm{Span}(u_\delta^*) \geq 1$ in the finite-sample analysis, where $\mathrm{Span}(v) := \max_{s\in\mathcal{S}} v(s) - \min_{s\in\mathcal{S}} v(s)$ for all $v \in \mathbb{R}^{|\mathcal{S}|}$. Note that the solution to (3.3) need not be unique. However, our main complexity result (Theorem 1 below) holds for any u_δ^* satisfying (3.3).

If the robust Bellman equation (3.3) admits a solution (u_δ^*, g), then $g_\delta^*(s) = g$ for every $s \in \mathcal{S}$, hence $g = g_\delta^*$. By Wang & Si (2025, Theorem 1), this property holds for $r(s,a)$; the same verification argument extends directly to our setting with bounded transition-dependent rewards $r(s,a,s')$.

4 Structural Properties and Learning Guarantees

This section presents our main structural and learning results. We first establish conditions for the robust average-reward Bellman equation, and then derive the learning algorithm and its finite-sample guarantees.

A key step in our finite-sample analysis is to bound the span of the optimal discounted value uniformly as γ approaches one. Chen et al. (2025) obtain such control through uniform ergodicity and mixing-time bounds. Our result instead uses a weaker sufficient condition: the existence of a constant-gain solution to the optimal average-reward Bellman equation. Specifically, a solution

(u_δ^*, g_δ^*) satisfying $\mathcal{T}^*(u_\delta^*) = u_\delta^* + g_\delta^* \mathbf{1}$ yields $\mathrm{Span}(v_\gamma^*) \le 2\,\mathrm{Span}(u_\delta^*)$ for every $\gamma \in (0,1)$ (Lemma 21). This comparison, building on Wang & Si (2025), allows us to control the discounted value span directly through the robust optimal bias span and express the resulting estimation bounds in terms of this quantity. We therefore first establish conditions ensuring a constant-gain Bellman solution for the uncertainty sets considered here, and then use this span comparison in the finite-sample analysis of Section 4.2.

4.1 Bellman optimality existence

In addressing different construction sources for uncertainty sets, this section presents two conclusions: for divergence-based uncertainty sets D_{KL} and D_{f_k}, when nominal kernel P satisfies weakly communicating and the uncertainty radius condition, then all kernels $Q \in \mathcal{P}$ within the SA-rect and S-rect sets possess the weakly communicating property; this guarantees the existence of an optimal Bellman equation and ensures the existence of a constant optimal robust gain. Meanwhile, since distance-based uncertainty sets D_{TV} and W_ℓ do not require absolute continuity, a positive uncertainty radius suffices to ensure the existence of a constant optimal gain function as a solution to the Bellman equation.

Definition 5. Define the minimum support as:

$$p_\wedge := \min_{(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}} \{p_{s,a}(s') : p_{s,a}(s') > 0\}. \tag{4.1}$$

Assumption 1. *The nominal MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$ is weakly communicating. Under either SA- or S-rectangularity, for $U \in \{D_{\mathrm{KL}}, D_{f_k}\}$, suppose the uncertainty radius satisfies*

$$\delta < \begin{cases} \frac{1}{\zeta} \log\left(\frac{1}{1-p_\wedge}\right), & U = D_{\mathrm{KL}}, \\ \frac{p_\wedge}{k\zeta}, & U = D_{f_k}. \end{cases} \quad \text{where} \quad \zeta := \begin{cases} 1, & \text{SA-rectangularity}, \\ |\mathcal{A}|, & \text{S-rectangularity}. \end{cases}$$

Inspired by Wang & Si's (2025) adversarial weak-communication framework, we derive explicit conditions on the nominal kernel P and radius δ that ensure every $Q \in \mathcal{P}$ is weakly communicating. We impose radius restrictions for KL- and f_k-divergence balls that preserve nominal transition supports, transferring nominal weak communication to every $Q \in \mathcal{P}$. Together with compactness and convexity, this establishes Bellman existence under both rectangularity structures through explicit nominal-model and radius conditions.

Proposition 1. *Under Assumption 1, the optimal average-reward Bellman equation (3.3) admits a solution (u_δ^*, g_δ^*) with a unique constant gain:*

$$\mathcal{T}^*(u_\delta^*) = u_\delta^* + g_\delta^* \mathbf{1}.$$

Assumption 2. *For distance-based uncertainty where $U \in \{d_{\mathrm{TV}}, W_\ell\}$, under both SA- and S-rectangularity suppose the uncertainty radius satisfies $\delta > 0$.*

Under Assumption 2, following Wang & Si (2025, Corollary 7.1), any radius $\delta > 0$ admits a feasible full-support perturbation. This yields a uniform bound on $\mathrm{Span}(v_\gamma^*)$ and, through a limit of normalized discounted values, a constant-gain Bellman solution for transition-dependent rewards under both rectangularity structures.

Proposition 2. *Under Assumption 2, the optimal average-reward Bellman equation (3.3) admits a solution (u_δ^*, g_δ^*) with a unique constant gain:*

$$\mathcal{T}^*(u_\delta^*) = u_\delta^* + g_\delta^* \mathbf{1}.$$

Corollary 1. *Under Assumption 1 or Assumption 2, there exists a stationary policy π^* attaining the robust optimal gain g_δ^*. In particular, $\pi^* \in \Pi_{\mathrm{SD}}$ under SA-rectangularity, while $\pi^* \in \Pi_{\mathrm{SR}}$ under S-rectangularity.*

Remark. Propositions 1–2 and Corollary 1 strengthen the structural theory in two directions. First, compared with Wang & Si (2025), we give explicit uncertainty-set-specific conditions ensuring a constant-gain Bellman solution. Second, under the same conditions, we establish stationary average-reward optimal policies for both SA- and S-rectangular uncertainty, providing explicit sufficient conditions for when S-rectangular average-reward optimality can be achieved by a stationary policy (Grand-Clément et al., 2026).

The proofs are given in Appendix A. These results yield uniform control of discounted value spans through the robust optimal bias span, providing the basis for the finite-sample guarantees in the next section without assuming uniform ergodicity.

4.2 Robust Average-Reward Learning

With the constant-gain Bellman solution established, we now study how to estimate the robust optimal average reward and learn a near-optimal policy from samples of the nominal transition kernel. Following the reduction approach of Chen et al. (2025), Algorithm 1 constructs an empirical robust discounted MDP with $\gamma_n = 1 - n^{-1/2}$ and solves its Bellman equation. It returns the empirical optimal policy $\widehat{\pi}^*$ and uses $(1-\gamma_n)\widehat{v}^*_{\gamma_n}$ to estimate $g^*_\delta \mathbf{1}$. We then establish finite-sample bounds for both the estimation error and the average-reward suboptimality of the returned policy under the true uncertainty set $\mathcal{P}$.

Generative model. We assume the accessability to a simulator, i.e., a *generative model* that allows the agent to sample independently from the nominal transition kernel $p_{s,a}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Given a n, we sample i.i.d. $\{S^{(1)}_{s,a}, \ldots, S^{(n)}_{s,a}\}$ from $p_{s,a}$ and construct the empirical kernel transition $\widehat{P} = \{\widehat{p}_{s,a}, (s,a) \in \mathcal{S} \times \mathcal{A}\}$ where

$$\widehat{p}_{s,a}(s') := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{S^{(i)}_{s,a} = s'\right\} \tag{4.2}$$

Algorithm 1 Distributionally Robust Average-Reward Learning (DR-ARL)

Input: sample size $n \geq 1$, $U \in \{D_{\mathrm{KL}}, D_{f_k}, d_{\mathrm{TV}}, W_\ell\}$, uncertainty radius δ.

Set $\gamma_n := 1 - n^{-1/2}$, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, construct an n-sample empirical transition kernel $\widehat{p}_{s,a}$ according to (4.2).

SA-rectangular: For each (s,a), set $\widehat{\mathcal{P}} = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{\mathcal{P}}_{s,a}(U,\delta)$ where $\widehat{\mathcal{P}}_{s,a}(U,\delta) := \{q_{s,a} \in \Delta(\mathcal{S}) : U(q_{s,a}, \widehat{p}_{s,a}) \leq \delta\}$,

Solve the empirical optimal discounted Bellman equation

$$\widehat{v}^*_{\gamma_n}(s) = \max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\widehat{\mathcal{P}}_{s,a}} \sum_{s'\in\mathcal{S}} q_{s,a}(s')\big(r(s,a,s') + \gamma_n \widehat{v}^*_{\gamma_n}(s')\big), \tag{4.3}$$

and for each $s \in \mathcal{S}$, let $\widehat{\pi}^*(s)$ be any action that maximize the right-hand side of (4.3).

S-rectangular: Construct $\widehat{\mathcal{P}} = \prod_{s\in\mathcal{S}} \widehat{\mathcal{P}}_s$ where $\widehat{\mathcal{P}}_s$ is jointly around $(\widehat{p}_{s,a})_{a\in\mathcal{A}}$ with radius δ.

Solve the empirical optimal discounted Bellman equation

$$\widehat{v}^*_{\gamma_n}(s) = \max_{\phi\in\Delta(\mathcal{A})} \inf_{q_s\in\widehat{\mathcal{P}}_s} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \phi(a) q_{s,a}(s')\big(r(s,a,s') + \gamma_n \widehat{v}^*_{\gamma_n}(s')\big), \tag{4.4}$$

and for each $s \in \mathcal{S}$, let $\widehat{\pi}^*(\cdot|s)$ be any action distribution that maximize the right-hand side of (4.4).

return $\widehat{\pi}^*$ and $\widehat{v}^*_{\gamma_n}/\sqrt{n}$.

We now state the gain-estimation and policy guarantees for Algorithm 1.

Theorem 1. *Suppose either Assumption 1 or Assumption 2 holds. Let (u^*_δ, g^*_δ) be any solution of the corresponding optimal average-reward Bellman equation (3.3). Fix $\beta \in (0,1)$. If*

$$n \geq \frac{16}{p_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right),$$

then the output of Algorithm 1 satisfies both of the following inequalities simultaneously with probability at least $1-\beta$:

$$\left\| \frac{\widehat{v}^*_{\gamma_n}}{\sqrt{n}} - g^*_\delta \mathbf{1} \right\|_\infty \leq \frac{16\,\mathrm{Span}(u^*_\delta)}{\sqrt{n p_\wedge}} \sqrt{\log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)} \tag{4.5}$$

$$\left\| g^*_\delta \mathbf{1} - g^{\widehat{\pi}^*}_\delta \right\|_\infty \leq \frac{151\,\mathrm{Span}(u^*_\delta)}{\sqrt{n p_\wedge}} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right) \tag{4.6}$$

where $g_\delta^{\widehat{\pi}^*}(s)$ is the robust average reward of the returned policy from initial state s, evaluated over the true uncertainty set $\mathcal{P}$ according to Section 3.3.

We develop a unified Bellman-operator perturbation analysis for discounted reduction, establishing finite-sample guarantees for both gain estimation and policy learning. Across all eight uncertainty configurations, our algorithm attains $\widetilde{O}(n^{-1/2})$ error bounds without prior knowledge of $p_\wedge$ or the robust optimal bias span.

Remark. Our upper bounds hold for any u_δ^* satisfying the robust Bellman equation (3.3). While Roch et al. (2026, Remark 3.3) claim that their results can be extended to weakly communicating settings, their complexity measure depends on the maximum of $\mathrm{Span}(u_\delta^*)$ over all solutions u_δ^* to the robust Bellman equation.

The next theorem makes the bias-span dependence in Theorem 1 explicit for all four uncertainty sets under both rectangularity structures.

Theorem 2. *Suppose Assumption 1 or Assumption 2 holds. Let $c = \max_{x,y\in\mathcal{S}} \rho(x,y)$. If*

$$\delta \le \begin{cases} \dfrac{p_\wedge}{8\zeta}, & U = D_{\mathrm{KL}}, \\ \dfrac{p_\wedge}{\zeta \max\{8, 4k\}}, & U = D_{f_k}, \\ 1, & U = d_{\mathrm{TV}}, \\ c, & U = W_\ell, \end{cases} \quad \textit{where} \quad \zeta := \begin{cases} 1, & \textit{SA-rectangularity}, \\ |\mathcal{A}|, & \textit{S-rectangularity}, \end{cases}$$

then the corresponding Bellman equation (3.3) admits a solution (u_δ^*, g_δ^*) satisfying

$$\mathrm{Span}(u_\delta^*) \le \begin{cases} \dfrac{2}{2-p_\wedge}\left[\left(\dfrac{2}{p_\wedge}\right)^{|\mathcal{S}|-1} - 1\right], & U \in \{D_{\mathrm{KL}}, D_{f_k}\}, \\ 1/\delta, & U = d_{\mathrm{TV}}, \\ (c/\delta)^\ell, & U = W_\ell. \end{cases}$$

Therefore, combining Theorems 1 and 2, the sample complexity for learning an ε-optimal policy is

$$\widetilde{O}\left(|\mathcal{S}||\mathcal{A}| p_\wedge^{-2} \varepsilon^{-2} \begin{cases} \left[(2/p_\wedge)^{|\mathcal{S}|-1} - 1\right]^2, & U \in \{D_{\mathrm{KL}}, D_{f_k}\}, \\ \delta^{-2}, & U = d_{\mathrm{TV}}, \\ (c/\delta)^{2\ell}, & U = W_\ell. \end{cases}\right).$$

The TV and Wasserstein upper bounds in Theorem 2 are tight, as shown by the two-state examples in the remark in Appendix F. Furthermore, for divergence-based uncertainty sets, Theorem 3 below shows that the exponential term $p_\wedge^{-|\mathcal{S}|}$ in the span bound is sharp.

Theorem 3. *For every $0 < p_\wedge \le 1/2$, there exists a weakly communicating nominal MDP such that, under either KL or f_k uncertainty, for every sufficiently small $\delta > 0$,*

$$\mathrm{Span}(u_\delta^*) \ge \frac{p_\wedge^{-(|\mathcal{S}|-1)} - 1}{1 - p_\wedge}.$$

The proof of Theorems 1, 2 and 3 are given in Appendix F.

5 Numerical Experiments

Theorem 1 shows that the gain-estimation error decreases at a rate of $n^{-1/2}$. We verify our main theorem with numerical validation. Considering both divergence-based and metric-based uncertainty sets under the two rectangularity structures. On the constructed MDP instance 6, we ran Algorithm 1 under the four uncertainty sets: KL-divergence, χ^2-divergence, TV and Wasserstein uncertainty sets under both SA- and S-rectangularity.

Definition 6. Consider the family of MDPs with state space $\mathcal{S} = \{0, 1, 2\}$ and action space $\mathcal{A} = \{a_1, a_2\}$. For $p \in (0, 1)$, the nominal transition rows, listed in state order $(0, 1, 2)$, are

$$p_{0,a_1} = (p, 1-p, 0), \qquad p_{0,a_2} = (p, 0, 1-p),$$
$$p_{1,a_1} = p_{2,a_2} = (0, 1, 0), \qquad p_{1,a_2} = p_{2,a_1} = (0, 0, 1).$$

For every $x \in \mathcal{S}$, the rewards are $r(0, a, x) = 0$ for $a \in \mathcal{A}$, and $r(s, a_1, x) = 1$, $r(s, a_2, x) = 5/6$ for $s \in \{1, 2\}$.

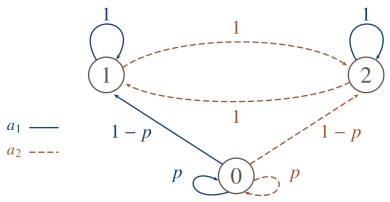
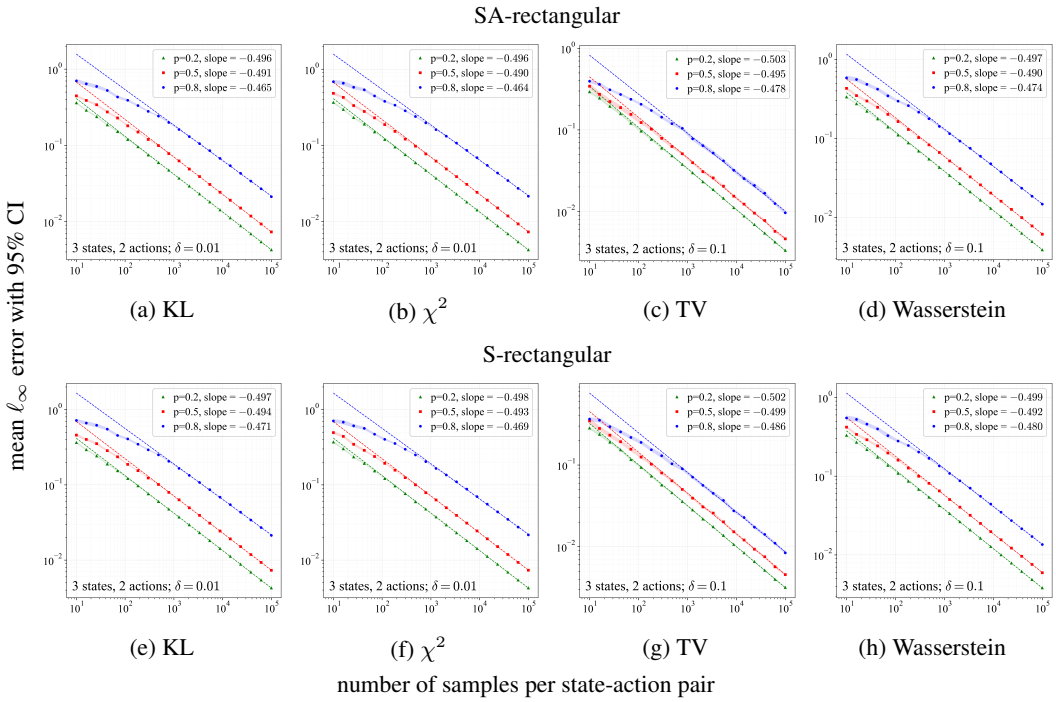


Figure 1: Nominal MDP in Definition 6.

This family, adapted from Wan et al. (2024, Figure 2(b)), is weakly communicating but not unichain: state 0 is transient under every policy, and choosing a_1 at states 1 and 2 creates two recurrent classes. We use $p \in \{0.2, 0.5, 0.8\}$, with $\delta = 0.01$ for KL and χ^2 ($k = 2$), and $\delta = 0.1$ for TV and Wasserstein. For Wasserstein, we set $\ell = 2$ and $\rho(s, s') = \mathbb{1}\{s \neq s'\}$. We evaluate 20 sample sizes from 10 to 10^5, with 30 independent trials per sample size. Figure 2 reports the results.



(a) KL (b) χ^2 (c) TV (d) Wasserstein

(e) KL (f) χ^2 (g) TV (h) Wasserstein

Figure 2: Mean ℓ_∞ gain-estimation error of Algorithm 1 versus the number of samples per state-action pair for $p \in \{0.2, 0.5, 0.8\}$. Shaded bands indicate percentile-bootstrap 95% confidence intervals over 30 trials. Dashed lines are log–log fits using $n \geq 10^3$, extended over the full plotted range.

Figure 2 plots the mean ℓ_∞ estimation error against the number of samples per state-action pair n, with both axes on logarithmic scales. Across all eight configurations and all three values of p, the fitted slopes range from -0.503 to -0.464, indicating an empirical error decay close to $n^{-1/2}$. This agrees with Theorem 1, which implies an $O(\epsilon^{-2})$ sample complexity, up to problem-dependent and logarithmic factors, for achieving estimation error at most ϵ. Additional large-scale experiments on a nominal MDP with 20 states and 30 actions are provided in Appendix G for all eight combinations of rectangularity structures and uncertainty sets, where the fitted slopes range from -0.527 to -0.490 and exhibit the same convergence behavior.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the design of numerical experiments and with drafting and polishing parts of the manuscript. The theoretical results and proofs were developed by the authors. We reviewed all AI-assisted writing to ensure that it accurately presents our arguments and results, and assessed suggestions for the experimental design before use. We take responsibility for the final content of this work, including text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work concerns theoretical analysis and numerical experiments on synthetic Markov decision processes. The experiments do not involve human participants or personal data.

REPRODUCIBILITY STATEMENT

The model, uncertainty sets, and assumptions are specified in Sections 3 and 4, and Algorithm 1 describes the learning procedure. Proofs of the theoretical results are provided in Appendices A–F. The experimental setups are described in Section 5 and Appendix G.

REFERENCES

Priyank Agrawal and Shipra Agrawal. Optimistic q-learning for average reward and episodic reinforcement learning extended abstract. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1–1. PMLR, 30 Jun–04 Jul 2025.

Zaiwei Chen. Non-asymptotic guarantees for average-reward q-learning with adaptive stepsizes. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 147194–147249. Curran Associates, Inc., 2025. doi: 10.52202/085713-4925.

Zijun Chen, Shengbo Wang, and Nian Si. Sample complexity of distributionally robust average-reward reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zijun Chen, Zaiwei Chen, Nian Si, and Shengbo Wang. Achieving ε^{-2} dependence for average-reward q-learning with a new contraction principle, 2026.

Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards minimax optimality of model-based robust reinforcement learning. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024a.

Pierre Clavier, Laixi Shi, Erwan Le Pennec, Eric Mazumdar, Adam Wierman, and Matthieu Geist. Near-optimal distributionally robust reinforcement learning with general l_p norms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1750–1810. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0056.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1578–1586. PMLR, 10–15 Jul 2018.

Julien Grand-Clément, Marek Petrik, and Nicolas Vieille. Beyond discounted returns: Robust Markov decision processes with average and Blackwell optimality. *Operations Research*, 74 (4):2088–2112, 2026.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3389–3396, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Yujia Jin and Aaron Sidford. Efficiently solving MDPs with stochastic mirror descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4890–4900, 2020.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward MDPs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5055–5064, 2021.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Yann Le Tallec. *Robust, risk-sensitive, and data-driven control of Markov decision processes*. PhD thesis, Massachusetts Institute of Technology, 2007.

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for MDPs via anchoring. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32907–32929, 2025.

Jongmin Lee, Ernest K. Ryu, and Vaneet Aggarwal. Learning policy from a single trajectory in average-reward Markov decision process. *arXiv preprint arXiv:2606.16729*, 2026.

Zhenghao Li, Shengbo Wang, and Nian Si. Near-optimal sample complexities of divergence-based S-rectangular distributionally robust reinforcement learning. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 937–945, 2026.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13623–13643, 2022.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9582–9602, 2022.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2005. Unabridged paperback reprint of the 1994 edition.

Zachary Roch and Yue Wang. Distributionally robust Markov games with average reward. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zachary Roch, George Atia, and Yue Wang. Model-free robust average-reward reinforcement learning with sample complexity analysis. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Operations Research*, 74(5):2438–2459, 2026.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

Yi Wan, Huizhen Yu, and Richard S. Sutton. On convergence of average-reward Q-learning in weakly communicating Markov decision processes. *arXiv preprint arXiv:2408.16262*, 2024.

Shengbo Wang and Nian Si. Bellman optimality of average-reward robust Markov decision processes with a constant gain. *arXiv preprint arXiv:2509.14203*, 2025.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust Q-learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3370–3398, 2023a.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward Markov decision processes. In *International Conference on Learning Representations*, 2024a.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust Q-learning. *Journal of Machine Learning Research*, 25(341): 1–77, 2024b.

Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15215–15223, 2023b.

Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36431–36469, 2023c.

Yue Wang, Alvaro Velasquez, George Atia, Ashley Prater-Bennette, and Shaofeng Zou. Robust average-reward reinforcement learning. *Journal of Artificial Intelligence Research*, 80:719–803, 2024c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Yang Xu, Washim Mondal, and Vaneet Aggarwal. Finite-sample analysis of policy evaluation for robust average reward reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 151870–151915, 2025.

Yang Xu, Swetha Ganesh, and Vaneet Aggarwal. Efficient Q-learning and actor–critic methods for robust average-reward reinforcement learning. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7613–7650, 2026.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9728–9754, 2023.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022.

Yuepeng Yang, Yuxin Chen, and Yuejie Chi. Robust average-reward Markov decision processes: Minimax-optimal learning via plug-in reductions. *arXiv preprint arXiv:2608.06545*, 2026.

Luhao Zhang, Jincheng Yang, and Rui Gao. A short and general duality proof for Wasserstein distributionally robust optimization. *Operations Research*, 73(4):2146–2155, 2025.

Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward TD learning and Q-learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1230–1242, 2021.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward Markov decision processes. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5476–5477, 2023.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33455–33504, 2024.

Matthew Zurek and Yudong Chen. Span-agnostic optimal sample complexity and oracle inequalities for average-reward RL. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 6156–6209, 2025a.

Matthew Zurek and Yudong Chen. The plug-in approach for average-reward and discounted MDPs: Optimal sample complexity analysis. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 1386–1387, 2025b.

A Basic Properties

This appendix provides results underlying Propositions 1 and 2 and proves Corollary 1. We first show that the inf–sup equation has a solution, justify exchanging the infimum and maximum, and prove uniqueness of the constant gain when the rewards depend on the realized next state. We then apply these results to KL- and f_k-divergence uncertainty sets and treat TV and Wasserstein uncertainty sets by taking a limit of normalized discounted values.

For $p \in \Delta(\mathcal{S})$ and $f : \mathcal{S} \to \mathbb{R}$, write $p[f] := \sum_{x \in \mathcal{S}} p(x) f(x)$.

Definition 7. The adversary is weakly communicating if, for every $Q \in \mathcal{P}$, the MDP with transition kernel Q is weakly communicating in the sense of Definition 1.

The preceding definition is identical to the adversary-side definition in Wang & Si (2025, Definition 3). The inf–sup equation in the following lemma is adapted from Wang & Si (2025, Theorem 6) to rewards that depend on the realized next state.

Lemma 1. *Suppose that $\mathcal{P} = \prod_{s \in \mathcal{S}} \mathcal{P}_s$ is S-rectangular, the adversary is weakly communicating, and every $\mathcal{P}_s$ is compact. Then there exists $(u', \alpha') \in \mathbb{R}^{\mathcal{S}} \times [0, 1]$ such that*

$$u'(s) = \inf_{q_s \in \mathcal{P}_s} \max_{\phi \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \phi(a) q_{s,a}(s') \big[r(s, a, s') - \alpha' + u'(s') \big], \qquad s \in \mathcal{S}. \tag{IS}$$

Proof. For $\pi \in \Pi_{\mathrm{SR}}$ and $Q \in \mathcal{P}$, write

$$V_\gamma^{\pi, Q}(s) := \mathrm{E}_Q^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t, S_{t+1}) \;\middle|\; S_0 = s \right].$$

Fix $Q = \{q_{s,a}\}_{s,a} \in \mathcal{P}$ and define

$$\bar{r}_Q(s, a) := \sum_{s' \in \mathcal{S}} q_{s,a}(s') r(s, a, s').$$

For every $\pi \in \Pi_{\mathrm{SR}}$, conditional expectation gives

$$\mathrm{E}_Q^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t, S_{t+1}) \;\middle|\; S_0 = s \right] = \mathrm{E}_Q^\pi \left[\sum_{t=0}^{\infty} \gamma^t \bar{r}_Q(S_t, A_t) \;\middle|\; S_0 = s \right].$$

Thus, once Q is fixed, this is an ordinary discounted MDP, and the stationary ε-optimality argument in the proof of Wang & Si (2025, Theorem 6) applies.

The finite-cover construction and Lemmas 5–8 in Appendix D of Wang & Si (2025) concern only induced transition kernels, communicating classes, and hitting times, and hence remain valid when the reward is $r(s, a, s')$. They yield a finite cover $\{G_{Q_k}\}_{k \in B'}$ of $\mathcal{P}$, classes C_{Q_k}, and $\delta' > 0$ such that, for $Q \in G_{Q_k}$ and $y \in C_{Q_k}$, some $\widetilde{\pi} \in \Pi_{\mathrm{SR}}$ satisfies

$$\max_{w \in \mathcal{S}} \mathrm{E}_Q^{\widetilde{\pi}}[\tau_y \mid S_0 = w] \le \frac{|\mathcal{S}|}{\delta'}.$$

Moreover, with

$$T_k := \inf\{t \ge 0 : S_t \in C_{Q_k}\},$$

Wang & Si (2025, Lemma 6) and the finite cover give

$$\max_{k \in B'} \sup_{\substack{Q \in G_{Q_k}, \pi \in \Pi_{\mathrm{SR}} \\ w \in \mathcal{S}}} \mathrm{E}_Q^\pi[T_k \mid S_0 = w] < \infty.$$

For $\gamma \in (0, 1)$, let v'_γ solve

$$v'_\gamma(s) = \inf_{q_s \in \mathcal{P}_s} \max_{\phi \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \phi(a) q_{s,a}(s') \big[r(s, a, s') + \gamma v'_\gamma(s') \big]. \tag{D}$$

Fix $\varepsilon > 0$. For each $s \in \mathcal{S}$, choose $q_{\varepsilon,s} = \{q_{\varepsilon,s,a}\}_{a \in \mathcal{A}} \in \mathcal{P}_s$ such that

$$\max_{\phi \in \Delta(\mathcal{A})} \sum_{a, s'} \phi(a) q_{\varepsilon,s,a}(s') \big[r(s, a, s') + \gamma v'_\gamma(s') \big] \le v'_\gamma(s) + (1 - \gamma)\varepsilon.$$

By S-rectangularity, the statewise choices form $Q_\varepsilon \in \mathcal{P}$. Discounted fixed-point comparison gives, simultaneously for all $s \in \mathcal{S}$,

$$v'_\gamma(s) \le \max_{\pi \in \Pi_{\mathrm{SR}}} V_\gamma^{\pi, Q_\varepsilon}(s) \le v'_\gamma(s) + \varepsilon.$$

The first inequality follows because Q_ε is feasible in the infimum in (D), while $(1-\gamma)\varepsilon$ accumulates to at most ε. For this finite discounted MDP, choose $\pi_\varepsilon \in \Pi_{\mathrm{SD}}$ that is optimal for every initial state, so that

$$\max_{\pi \in \Pi_{\mathrm{SR}}} V_\gamma^{\pi, Q_\varepsilon}(s) = V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(s), \qquad s \in \mathcal{S}.$$

Choose $k \in B'$ with $Q_\varepsilon \in G_{Q_k}$ and let

$$\begin{aligned} s_\vee \in \arg\max_{s \in \mathcal{S}} v'_\gamma(s), \qquad s_\wedge \in \arg\min_{s \in \mathcal{S}} v'_\gamma(s), \\ y_\vee \in \arg\max_{y \in C_{Q_k}} V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y). \end{aligned} \tag{Y}$$

The preceding comparison gives

$$\begin{aligned} \mathrm{Span}(v'_\gamma) &\le V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(s_\vee) - V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(s_\wedge) + 2\varepsilon \\ &= \underbrace{V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(s_\vee) - V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y_\vee)}_{\xi_2} + \underbrace{V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y_\vee) - V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(s_\wedge)}_{\xi_1} + 2\varepsilon. \end{aligned} \tag{1}$$

Upper-bounding ξ_1. Fix $y \in C_{Q_k}$. By Wang & Si (2025, Lemma 8), there exists $\widetilde{\pi} \in \Pi_{\mathrm{SR}}$ such that

$$\max_{w \in \mathcal{S}} \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}[\tau_y \mid S_0 = w] \le \frac{|\mathcal{S}|}{\delta'}.$$

Since $V_\gamma^{\pi_\varepsilon, Q_\varepsilon}$ is the optimal discounted value for Q_ε, its Bellman inequality under $\widetilde{\pi}$ gives, by iteration up to τ_y, for every $x \in \mathcal{S}$,

$$\begin{aligned} V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(x) &\ge \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}\left[\sum_{t=0}^{\tau_y - 1} \gamma^t r(S_t, A_t, S_{t+1}) \;\middle|\; S_0 = x\right] \\ &\quad + \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}\left[\gamma^{\tau_y} V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y) \;\middle|\; S_0 = x\right] \\ &\ge V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y) \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}[\gamma^{\tau_y} \mid S_0 = x]. \end{aligned} \tag{2}$$

The stopped inequality follows by truncation and boundedness of the discounted value. The first sum includes the reward on the transition entering y. Hence

$$\begin{aligned} &V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y) - V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(x) \\ &\quad \le V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y) \left(1 - \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}[\gamma^{\tau_y} \mid S_0 = x]\right) \\ &\quad = (1-\gamma) V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y)\, \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}\left[\frac{1-\gamma^{\tau_y}}{1-\gamma} \;\middle|\; S_0 = x\right] \\ &\quad \le \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}[\tau_y \mid S_0 = x]. \end{aligned}$$

The last inequality follows from $0 \le V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y) \le (1-\gamma)^{-1}$ and $(1-\gamma^t)/(1-\gamma) = \sum_{j=0}^{t-1} \gamma^j \le t$. Taking $x = s_\wedge$ and $y = y_\vee$ therefore gives

$$\xi_1 \le \mathrm{E}_{Q_\varepsilon}^{\widetilde{\pi}}[\tau_{y_\vee} \mid S_0 = s_\wedge] \le \frac{|\mathcal{S}|}{\delta'}. \tag{3}$$

Upper-bounding ξ_2. Let $T_k := \inf\{t \geq 0 : S_t \in C_{Q_k}\}$. Then

$$\begin{aligned}
\xi_2 &= V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(s_\vee) - V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y_\vee) \\
&= \mathrm{E}_{Q_\varepsilon}^{\pi_\varepsilon}\left[\sum_{t=0}^{T_k-1} \gamma^t r(S_t, A_t, S_{t+1}) + \sum_{t=T_k}^{\infty} \gamma^t r(S_t, A_t, S_{t+1}) \;\middle|\; S_0 = s_\vee\right] \\
&\quad - V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y_\vee) \\
&\overset{\text{(i)}}{=} \mathrm{E}_{Q_\varepsilon}^{\pi_\varepsilon}\left[\sum_{t=0}^{T_k-1} \gamma^t r(S_t, A_t, S_{t+1}) + \gamma^{T_k} V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(S_{T_k}) \;\middle|\; S_0 = s_\vee\right] \\
&\quad - V_\gamma^{\pi_\varepsilon, Q_\varepsilon}(y_\vee) \\
&\overset{\text{(ii)}}{\leq} \mathrm{E}_{Q_\varepsilon}^{\pi_\varepsilon}\left[\sum_{t=0}^{T_k-1} \gamma^t r(S_t, A_t, S_{t+1}) \;\middle|\; S_0 = s_\vee\right] \\
&\leq \mathrm{E}_{Q_\varepsilon}^{\pi_\varepsilon}[T_k \mid S_0 = s_\vee].
\end{aligned} \tag{4}$$

Here *(i)* follows from the strong Markov property, while *(ii)* follows from $S_{T_k} \in C_{Q_k}$, the choice of $y_\vee$ in (Y), and nonnegativity of the discounted value. Consequently,

$$\xi_2 \leq \max_{k \in B'} \sup_{\substack{Q \in G_{Q_k},\, \pi \in \Pi_{\mathrm{SR}} \\ w \in \mathcal{S}}} \mathrm{E}_Q^\pi[T_k \mid S_0 = w] < \infty. \tag{5}$$

Combining (1)–(5), letting $\varepsilon \downarrow 0$, and applying the uniform entrance-time bound yields

$$\sup_{\gamma \in (0,1)} \mathrm{Span}(v'_\gamma) \leq \frac{|\mathcal{S}|}{\delta'} + \max_{k \in B'} \sup_{\substack{Q \in G_{Q_k},\, \pi \in \Pi_{\mathrm{SR}} \\ w \in \mathcal{S}}} \mathrm{E}_Q^\pi[T_k \mid S_0 = w] < \infty. \tag{6}$$

Fix $s_0 \in \mathcal{S}$ and set

$$h_\gamma := v'_\gamma - v'_\gamma(s_0)\mathbf{1}, \qquad \alpha_\gamma := (1-\gamma)v'_\gamma(s_0).$$

Since $r(s,a,s') \in [0,1]$, (D) gives $0 \leq v'_\gamma(s) \leq (1-\gamma)^{-1}$ and hence $0 \leq \alpha_\gamma \leq 1$. By (6), along some $\gamma_n \uparrow 1$,

$$h_{\gamma_n} \to u', \qquad \alpha_{\gamma_n} \to \alpha' \in [0,1].$$

The normalized form of (D) is

$$h_\gamma(s) = \inf_{q_s \in \mathcal{P}_s} \max_{\phi \in \Delta(\mathcal{A})} \sum_{a,s'} \phi(a) q_{s,a}(s') \big[r(s,a,s') - \alpha_\gamma + \gamma h_\gamma(s')\big].$$

Moreover,

$$\begin{aligned}
&\max_{s \in \mathcal{S}} \sup_{\substack{\phi \in \Delta(\mathcal{A}) \\ q_s \in \mathcal{P}_s}} \left| \sum_{a,s'} \phi(a) q_{s,a}(s') \big[r(s,a,s') - \alpha_{\gamma_n} + \gamma_n h_{\gamma_n}(s')\big] \right. \\
&\qquad \left. - \sum_{a,s'} \phi(a) q_{s,a}(s') \big[r(s,a,s') - \alpha' + u'(s')\big] \right| \\
&\leq |\alpha_{\gamma_n} - \alpha'| + \max_{s' \in \mathcal{S}} |\gamma_n h_{\gamma_n}(s') - u'(s')| \longrightarrow 0.
\end{aligned}$$

The normalization-and-subsequence argument in Wang & Si (2025, Theorem 4) now allows passage to the limit in (D), giving (IS). □

The following lemma adapts Wang & Si (2025, Theorem 7(1)) to rewards that depend on the realized next state.

Lemma 2. *Suppose that (u', α') solves the inf–sup equation in Lemma 1. Assume that the controller action set is compact and convex and that every $\mathcal{P}_s$ is convex. Then (u', α') satisfies the optimal average-reward Bellman equation (3.3) defined in the preliminaries.*

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

$$F_s(\phi, q_s) := \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \phi(a) q_{s,a}(s') \big[r(s,a,s') - \alpha' + u'(s')\big].$$

The function F_s is linear and continuous in each argument. Since $\mathcal{A}$ is finite, the controller action set $\Delta(\mathcal{A})$ is compact and convex; together with the convexity of $\mathcal{P}_s$, this satisfies the conditions of Sion's minimax theorem (Sion, 1958, Corollary 3.3). Hence,

$$\inf_{q_s \in \mathcal{P}_s} \max_{\phi \in \Delta(\mathcal{A})} F_s(\phi, q_s) = \max_{\phi \in \Delta(\mathcal{A})} \inf_{q_s \in \mathcal{P}_s} F_s(\phi, q_s).$$

Combining this equality with the inf–sup equation in Lemma 1 gives the claimed Bellman equation (3.3). □

Below, we show that if the optimal average-reward Bellman equation (3.3) admits a solution, then the constant gain must be unique.

Lemma 3. *For either the SA-rectangular or S-rectangular optimal average-reward Bellman operator $\mathcal{T}^*$ in (3.2), suppose that (u, α) and (v, β) are two solutions of (3.3):*

$$\mathcal{T}^*(u) = u + \alpha \mathbf{1}, \qquad \mathcal{T}^*(v) = v + \beta \mathbf{1}.$$

Then $\alpha = \beta$. Thus only the constant gain is necessarily unique; the relative value function need not be unique.

Proof. For both the SA-rectangular and S-rectangular operators in (3.2),

$$u \le v \quad \Longrightarrow \quad \mathcal{T}^*(u) \le \mathcal{T}^*(v), \qquad \mathcal{T}^*(v + c\mathbf{1}) = \mathcal{T}^*(v) + c\mathbf{1}, \quad c \in \mathbb{R}.$$

The term $r(s, a, s')$ is unchanged when the relative value function is compared or shifted, so it does not affect these properties.

Let

$$c := \max_{s \in \mathcal{S}} \{u(s) - v(s)\}, \qquad \bar{s} \in \arg\max_{s \in \mathcal{S}} \{u(s) - v(s)\}.$$

Then $u \le v + c\mathbf{1}$ and $u(\bar{s}) = v(\bar{s}) + c$. Monotonicity and additive homogeneity give

$$u + \alpha \mathbf{1} = \mathcal{T}^*(u) \le \mathcal{T}^*(v + c\mathbf{1}) = v + (\beta + c)\mathbf{1}.$$

Evaluating at $\bar{s}$ yields $\alpha \le \beta$. Interchanging the two solutions gives $\beta \le \alpha$, and therefore $\alpha = \beta$. □

Lemma 4. *Under Assumption 1, every adversarial transition kernel $Q \in \mathcal{P}$ is weakly communicating.*

Proof. Fix $Q \in \mathcal{P}$. For every (s, a), the definition of the divergence-based uncertainty sets gives $q_{s,a} \ll p_{s,a}$. Thus,

$$p_{s,a}(s') = 0 \quad \Longrightarrow \quad q_{s,a}(s') = 0.$$

Conversely, let $p_{s,a}(s') > 0$ and suppose that $q_{s,a}(s') = 0$. In the KL case, applying the log-sum inequality to $\mathcal{S} \setminus \{s'\}$ gives

$$\begin{aligned} D_{\mathrm{KL}}(q_{s,a} \| p_{s,a}) &= q_{s,a}(s') \log\left(\frac{q_{s,a}(s')}{p_{s,a}(s')}\right) + \sum_{x \neq s'} q_{s,a}(x) \log\left(\frac{q_{s,a}(x)}{p_{s,a}(x)}\right) \\ &\ge q_{s,a}(s') \log\left(\frac{q_{s,a}(s')}{p_{s,a}(s')}\right) + \big(1 - q_{s,a}(s')\big) \log\left(\frac{1 - q_{s,a}(s')}{1 - p_{s,a}(s')}\right) \\ &= \log\left(\frac{1}{1 - p_{s,a}(s')}\right) \ge \log\left(\frac{1}{1 - p_\wedge}\right). \end{aligned}$$

Consequently, in the SA-rectangular case,

$$\delta \ge D_{\mathrm{KL}}(q_{s,a} \| p_{s,a}) \ge \log\left(\frac{1}{1 - p_\wedge}\right),$$

whereas in the S-rectangular case,

$$|\mathcal{A}|\delta \ge \sum_{b \in \mathcal{A}} D_{\mathrm{KL}}(q_{s,b} \| p_{s,b}) \ge D_{\mathrm{KL}}(q_{s,a} \| p_{s,a}) \ge \log\left(\frac{1}{1 - p_\wedge}\right).$$

Both conclusions contradict the corresponding KL radius condition in Assumption 1.

In the f_k-divergence case, convexity of $t \mapsto t^k$ gives

$$t^k \geq 1 + k(t-1), \qquad f_k(t) = \frac{t^k - kt + k - 1}{k(k-1)} \geq 0, \qquad t \geq 0,$$

and $f_k(0) = 1/k$. Therefore,

$$\begin{aligned} D_{f_k}(q_{s,a} \| p_{s,a}) &= \sum_{x \in \mathcal{S}} p_{s,a}(x) f_k\left(\frac{q_{s,a}(x)}{p_{s,a}(x)}\right) \\ &\geq p_{s,a}(s') f_k\left(\frac{q_{s,a}(s')}{p_{s,a}(s')}\right) = \frac{p_{s,a}(s')}{k} \geq \frac{p_\wedge}{k}. \end{aligned}$$

Thus, in the SA-rectangular case,

$$\delta \geq D_{f_k}(q_{s,a} \| p_{s,a}) \geq \frac{p_\wedge}{k},$$

while in the S-rectangular case,

$$|\mathcal{A}|\delta \geq \sum_{b \in \mathcal{A}} D_{f_k}(q_{s,b} \| p_{s,b}) \geq D_{f_k}(q_{s,a} \| p_{s,a}) \geq \frac{p_\wedge}{k}.$$

These conclusions contradict the corresponding f_k-divergence radius condition in Assumption 1. Hence

$$\operatorname{supp}(q_{s,a}) = \operatorname{supp}(p_{s,a}), \qquad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Therefore, for every $\pi \in \Pi_{\mathrm{SR}}$,

$$\sum_{a \in \mathcal{A}} \pi(a \mid s) p_{s,a}(s') > 0 \iff \sum_{a \in \mathcal{A}} \pi(a \mid s) q_{s,a}(s') > 0, \qquad s, s' \in \mathcal{S}.$$

Thus the transition kernels induced by P and Q under π have the same directed graph. The partition witnessing weak communication of the nominal MDP also witnesses weak communication for the MDP with transition kernel Q. Since $Q \in \mathcal{P}$ was arbitrary, the adversary is weakly communicating according to Definition 7. □

Proof of Proposition 1. By Lemma 4 and Definition 7, the adversary is weakly communicating. Under Assumption 1, the constraint $q_{s,a} \ll p_{s,a}$ defines a closed face of $\Delta(\mathcal{S})$. On this face, $D_{\mathrm{KL}}(q_{s,a} \| p_{s,a})$ is continuous and convex in $q_{s,a}$. Likewise, $D_{f_k}(q_{s,a} \| p_{s,a})$ is continuous and convex because f_k is continuous and convex. Hence every KL- or f_k-divergence set $\mathcal{P}_{s,a}(U, \delta)$, with $U \in \{D_{\mathrm{KL}}, D_{f_k}\}$, is a closed convex subset of the compact simplex $\Delta(\mathcal{S})$, and is therefore compact and convex.

In the SA-rectangular case, set $\mathcal{P}_s := \prod_{a \in \mathcal{A}} \mathcal{P}_{s,a}(U, \delta)$; these statewise products are compact and convex. In the S-rectangular case, the two defining functions

$$\sum_{a \in \mathcal{A}} D_{\mathrm{KL}}(q_{s,a} \| p_{s,a}) \quad \text{and} \quad \sum_{a \in \mathcal{A}} D_{f_k}(q_{s,a} \| p_{s,a})$$

are continuous and convex on a finite product of the same closed support faces. Their sublevel sets $\mathcal{P}_s(U, \delta)$ are therefore compact and convex. Thus, under either structure, $\mathcal{P} = \prod_{s \in \mathcal{S}} \mathcal{P}_s$ has compact and convex statewise sets. Finally, $\Delta(\mathcal{A})$ is compact and convex. Lemma 1 therefore yields a solution (u', α') of the inf–sup constant-gain equation, and Lemma 2 shows that the same pair solves the optimal average-reward Bellman equation (3.3). For the SA-rectangular model,

$$\begin{aligned} &\max_{\phi \in \Delta(\mathcal{A})} \inf_{q_s \in \prod_{a \in \mathcal{A}} \mathcal{P}_{s,a}} \sum_{a \in \mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + u')] \\ &\quad = \max_{a \in \mathcal{A}} \inf_{q_{s,a} \in \mathcal{P}_{s,a}} q_{s,a}[r(s,a,\cdot) + u']. \end{aligned}$$

Thus this is precisely the SA-rectangular operator in Definition 4. Taking $u^*_\delta = u'$ and $g^*_\delta = \alpha'$ establishes existence. By Lemma 3, every other solution has the same constant gain g^*_δ, which proves uniqueness. □

Proof of Proposition 2. The argument is the same for the SA-rectangular and S-rectangular formulations.

Choose any full-support $\nu \in \Delta(\mathcal{S})$ and, for $\varepsilon \in (0,1)$, set $\bar{Q} = \{\bar{q}_{s,a}\}_{s,a}$, where

$$\bar{q}_{s,a} := (1-\varepsilon)p_{s,a} + \varepsilon\nu.$$

For total variation,

$$d_{\mathrm{TV}}(\bar{q}_{s,a}, p_{s,a}) = \varepsilon d_{\mathrm{TV}}(\nu, p_{s,a}) \longrightarrow 0,$$

whereas, on the finite metric space $(\mathcal{S}, \rho)$,

$$W_\ell(\bar{q}_{s,a}, p_{s,a})^\ell \le \varepsilon \max_{x,y\in\mathcal{S}} \rho(x,y)^\ell \longrightarrow 0.$$

Both distances, as well as their finite action-wise sums, vanish as $\varepsilon \downarrow 0$. Thus, under either rectangularity structure, a common sufficiently small $\varepsilon > 0$ gives $\bar{Q} \in \mathcal{P}$. Moreover,

$$\bar{q}_\wedge := \min_{s,a,s'} \bar{q}_{s,a}(s') \ge \varepsilon \min_{s'} \nu(s') > 0.$$

Fix $\gamma \in (0,1)$ and choose $s_- \in \arg\min_x v^*_\gamma(x)$. For every (s,a),

$$\begin{aligned}\bar{q}_{s,a}[v^*_\gamma] &\le \bar{q}_{s,a}(s_-) \min_x v^*_\gamma(x) + \big(1 - \bar{q}_{s,a}(s_-)\big) \max_x v^*_\gamma(x) \\ &\le \bar{q}_\wedge \min_x v^*_\gamma(x) + (1-\bar{q}_\wedge) \max_x v^*_\gamma(x).\end{aligned}$$

Evaluating the inner infimum in $\mathcal{T}^*_\gamma$ from (3.1) at the feasible kernel $\bar{Q}$ gives, for both rectangularity structures,

$$\begin{aligned}\max_s v^*_\gamma(s) &\le \max_{s,a,s'} r(s,a,s') + \gamma\left(\bar{q}_\wedge \min_x v^*_\gamma(x) + (1-\bar{q}_\wedge)\max_x v^*_\gamma(x)\right), \\ \min_s v^*_\gamma(s) &\ge \min_{s,a,s'} r(s,a,s') + \gamma \min_x v^*_\gamma(x).\end{aligned}$$

Subtracting the two inequalities and using $v^*_\gamma = \mathcal{T}^*_\gamma(v^*_\gamma)$ yields

$$\begin{aligned}\mathrm{Span}(v^*_\gamma) &\le \max_{s,a,s'} r(s,a,s') - \min_{s,a,s'} r(s,a,s') \\ &\quad + \gamma(1-\bar{q}_\wedge)\,\mathrm{Span}(v^*_\gamma) \\ &\le 1 + \gamma(1-\bar{q}_\wedge)\,\mathrm{Span}(v^*_\gamma).\end{aligned}$$

and therefore

$$\mathrm{Span}(v^*_\gamma) \le \frac{1}{1-\gamma(1-\bar{q}_\wedge)} \le \frac{1}{\bar{q}_\wedge}, \qquad \gamma \in (0,1).$$

The minimizing kernel need not have full support: feasibility of $\bar{Q}$ alone is sufficient because the inner infimum is no larger than its value at $\bar{Q}$.

Fix $s_0 \in \mathcal{S}$. The uniform span bound gives

$$\left\|v^*_\gamma - v^*_\gamma(s_0)\mathbf{1}\right\|_\infty \le \frac{1}{\bar{q}_\wedge}.$$

Boundedness of the reward and the discounted Bellman equation for the operator in (3.1) also give

$$0 \le v^*_\gamma(s) \le \frac{1}{1-\gamma}, \qquad s \in \mathcal{S}.$$

Hence finite dimensionality yields a sequence $\gamma_n \uparrow 1$ such that

$$v^*_{\gamma_n} - v^*_{\gamma_n}(s_0)\mathbf{1} \longrightarrow u^*_\delta, \qquad (1-\gamma_n)v^*_{\gamma_n}(s_0) \longrightarrow g^*_\delta.$$

For the S-rectangular model, the normalized discounted Bellman equation using the operator (3.1) is

$$\begin{aligned}v^*_\gamma(s) - v^*_\gamma(s_0) = \max_{\phi\in\Delta(\mathcal{A})} \inf_{q_s\in\mathcal{P}_s} &\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \phi(a) q_{s,a}(s') \\ &\times \Big[r(s,a,s') - (1-\gamma)v^*_\gamma(s_0) + \gamma\big(v^*_\gamma(s') - v^*_\gamma(s_0)\big)\Big].\end{aligned}$$

For the SA-rectangular model, using the reduction in Definition 4, the right-hand side is replaced by

$$\max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\mathcal{P}_{s,a}} \sum_{s'\in\mathcal{S}} q_{s,a}(s')\Big[r(s,a,s') - (1-\gamma)v^*_\gamma(s_0) + \gamma\big(v^*_\gamma(s') - v^*_\gamma(s_0)\big)\Big].$$

The corresponding one-stage expressions converge uniformly because

$$\big|(1-\gamma_n)v^*_{\gamma_n}(s_0) - g^*_\delta\big| + \big\|\gamma_n\big(v^*_{\gamma_n} - v^*_{\gamma_n}(s_0)\mathbf{1}\big) - u^*_\delta\big\|_\infty \longrightarrow 0.$$

Since a max–infimum is 1-Lipschitz with respect to the uniform norm of its objective, passage to the limit in the corresponding equation gives (3.3):

$$\mathcal{T}^*(u^*_\delta) = u^*_\delta + g^*_\delta\mathbf{1}$$

under either rectangularity structure. Lemma 3 then implies that every other solution has the same constant gain g^*_δ. □

Proof of Corollary 1. Propositions 1 and 2 provide a solution (u^*_δ, g^*_δ) of (3.3). By Wang & Si (2025, Theorem 3 and Remark 3), a stationary policy attaining the maximization in the constant-gain Bellman equation is robust optimal. Their result is stated for rewards $r(s,a)$, and the same verification argument applies to our bounded transition-dependent rewards $r(s,a,s')$.

For the SA-rectangular case,

$$u^*_\delta(s) + g^*_\delta = \max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\mathcal{P}_{s,a}} q_{s,a}[r(s,a,\cdot) + u^*_\delta].$$

Since $\mathcal{A}$ is finite, for each $s\in\mathcal{S}$ there exists

$$a^*(s) \in \arg\max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\mathcal{P}_{s,a}} q_{s,a}[r(s,a,\cdot) + u^*_\delta].$$

Define $\pi^*(a \mid s) = \mathbf{1}\{a = a^*(s)\}$. Then π^* is stationary deterministic, so $\pi^*\in\Pi_{\mathrm{SD}}$, and it attains g^*_δ.

For the S-rectangular case,

$$u^*_\delta(s) + g^*_\delta = \max_{\phi\in\Delta(\mathcal{A})} \inf_{q_s\in\mathcal{P}_s} \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + u^*_\delta)]\,.$$

The objective is continuous in ϕ, and $\Delta(\mathcal{A})$ is compact, so a maximizer exists. Choose

$$\phi^*(s) \in \arg\max_{\phi\in\Delta(\mathcal{A})} \inf_{q_s\in\mathcal{P}_s} \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + u^*_\delta)]\,,$$

and define $\pi^*(\cdot \mid s) = \phi^*(s)$. Then π^* is stationary randomized, so $\pi^*\in\Pi_{\mathrm{SR}}$, and it attains g^*_δ.

To verify optimality for the kernel-sequence criterion, summing the Bellman inequality shows that either constructed policy has expected T-step reward at least $Tg^*_\delta - \mathrm{Span}(u^*_\delta)$ from every initial state under every $\mathbf{Q}\in(\mathcal{P})^{\mathbb{N}}$. Conversely, compactness and actionwise minimization under SA-rectangularity, or compactness, convexity, and the minimax argument in Lemma 2 under S-rectangularity, yield a fixed $Q = (q_{s,a})_{s,a}\in\mathcal{P}$ such that $q_{s,a}[r(s,a,\cdot) + u^*_\delta] \le u^*_\delta(s) + g^*_\delta$ for all (s,a). Against the constant sequence $(Q, Q, \ldots)$, summing this inequality bounds the expected T-step reward of every $\pi\in\Pi_{\mathrm{HD}}$ by $Tg^*_\delta + \mathrm{Span}(u^*_\delta)$. Dividing by T and taking $\limsup$ therefore proves that π^* attains the robust optimal gain for every initial state. □

B Notations and Dual Forms

Our algorithms replace the nominal transition distributions with estimates obtained from samples. We describe the resulting empirical DR-MDP and its Bellman operators, then introduce the event on which these estimates remain close to their nominal counterparts. We then state the fixed-policy dual representations and give their proofs.

For each $(s,a)\in\mathcal{S}\times\mathcal{A}$, let $S^{(1)}_{s,a}, \ldots, S^{(n)}_{s,a}$ be independent samples drawn from $p_{s,a}$. For a sample outcome ω, define the empirical transition distribution by

$$\widehat{p}^{\,\omega}_{s,a}(s') := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{S^{(i)}_{s,a}(\omega) = s'\}, \qquad s'\in\mathcal{S}.$$

For the observed samples, we suppress ω and write $\widehat{p}_{s,a}$. Since the samples are drawn from $p_{s,a}$, with probability one, $\mathrm{supp}(\widehat{p}_{s,a}) \subseteq \mathrm{supp}(p_{s,a})$ holds simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Replacing each nominal center $p_{s,a}$ in the uncertainty sets introduced in the preliminaries by $\widehat{p}_{s,a}$ gives the empirical DR-MDP. The uncertainty radius δ and the rectangularity structure are unchanged, so the empirical Bellman operators describe the same local controller–adversary games with uncertainty centered at the estimated transition distributions.

SA-rectangular model. The empirical uncertainty set factors over state-action pairs:

$$\widehat{\mathcal{P}} = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{\mathcal{P}}_{s,a}.$$

The adversary again minimizes each action's payoff independently. Replacing the uncertainty sets in the optimal discounted Bellman operator in Definition 4 gives

$$\widehat{\mathcal{T}}^*_\gamma(v)(s) := \max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\widehat{\mathcal{P}}_{s,a}} q_{s,a}[r(s,a,\cdot) + \gamma v].$$

The empirical optimal discounted value $\widehat{v}^*_\gamma$ is its fixed point:

$$\widehat{v}^*_\gamma = \widehat{\mathcal{T}}^*_\gamma(\widehat{v}^*_\gamma). \tag{B.1}$$

Removing the discount factor gives the corresponding empirical average-reward Bellman operator, as in Definition 4:

$$\widehat{\mathcal{T}}^*(v)(s) := \max_{a\in\mathcal{A}} \inf_{q_{s,a}\in\widehat{\mathcal{P}}_{s,a}} q_{s,a}[r(s,a,\cdot) + v].$$

S-rectangular model. Here the empirical uncertainty set factors over states:

$$\widehat{\mathcal{P}} = \prod_{s\in\mathcal{S}} \widehat{\mathcal{P}}_s.$$

The action-indexed transition distributions remain jointly constrained at each state. For a controller's action distribution $\phi \in \Delta(\mathcal{A})$, the adversary's response is described by the empirical discounted Bellman operator

$$\widehat{\mathcal{T}}^\phi_\gamma(v)(s) := \inf_{q_s\in\widehat{\mathcal{P}}_s} \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + \gamma v)].$$

Maximizing over ϕ gives the empirical counterpart of the optimal discounted Bellman operator (3.1):

$$\widehat{\mathcal{T}}^*_\gamma(v)(s) := \max_{\phi\in\Delta(\mathcal{A})} \widehat{\mathcal{T}}^\phi_\gamma(v)(s).$$

Its fixed point again determines the empirical optimal discounted value:

$$\widehat{v}^*_\gamma = \widehat{\mathcal{T}}^*_\gamma(\widehat{v}^*_\gamma). \tag{B.2}$$

The same local game without discounting gives the empirical average-reward Bellman operator, as in (3.2):

$$\widehat{\mathcal{T}}^*(v)(s) := \max_{\phi\in\Delta(\mathcal{A})} \inf_{q_s\in\widehat{\mathcal{P}}_s} \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + v)].$$

Fixed-policy discounted Bellman operators. For $\pi \in \Pi_{\mathrm{SR}}$ and $\gamma \in (0,1)$, define $\mathcal{T}^\pi_\gamma : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ under SA-rectangularity by

$$\mathcal{T}^\pi_\gamma(v)(s) := \sum_{a\in\mathcal{A}} \pi(a|s) \inf_{q_{s,a}\in\mathcal{P}_{s,a}} \sum_{s'\in\mathcal{S}} q_{s,a}(s')\big(r(s,a,s') + \gamma v(s')\big), \tag{B.3}$$

and under S-rectangularity by

$$\mathcal{T}^\pi_\gamma(v)(s) := \inf_{q_s\in\mathcal{P}_s} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \pi(a \mid s) q_{s,a}(s')\big(r(s,a,s') + \gamma v(s')\big). \tag{B.4}$$

The empirical fixed-policy discounted operator $\widehat{\mathcal{T}}^\pi_\gamma$ is obtained from $\mathcal{T}^\pi_\gamma$ by replacing the true uncertainty sets with their empirical counterparts while keeping π unchanged. Under either structure, $\mathcal{T}^*_\gamma(v) = \max_{\pi\in\Pi_{\mathrm{SR}}} \mathcal{T}^\pi_\gamma(v)$ componentwise, and the same identity holds for the empirical operators.

Under either rectangularity structure, comparing the empirical and true Bellman operators requires controlling the error in their centers. We therefore introduce the good event $\Omega_{n,\eta}$, on which all empirical transition probabilities are uniformly close to their nominal counterparts in relative error.

Definition 8. For $\eta > 0$, define the good event on which every empirical transition distribution remains close to its nominal transition distribution in relative error by

$$\Omega_{n,\eta} := \bigcap_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left\{ \omega : \left\| \frac{\widehat{p}^{\,\omega}_{s,a} - p_{s,a}}{p_{s,a}} \right\|_{L^\infty(p_{s,a})} \le \eta \right\},$$

where

$$\left\| \frac{\widehat{p}^{\,\omega}_{s,a} - p_{s,a}}{p_{s,a}} \right\|_{L^\infty(p_{s,a})} := \max_{s'\in\mathrm{supp}(p_{s,a})} \left| \frac{\widehat{p}^{\,\omega}_{s,a}(s') - p_{s,a}(s')}{p_{s,a}(s')} \right|.$$

Proposition 3. *For any $\beta \in (0,1)$, choose*

$$\eta = \frac{1}{3np_\wedge} \log \frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta} + \sqrt{\frac{2}{np_\wedge} \log \frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}}.$$

Then the good event satisfies

$$\mathbb{P}\big(\Omega^c_{n,\eta}\big) \le \beta.$$

This proposition is the same as Chen et al. (2025, Lemma A.2), written in the present notation.

B.1 Uncertainty Sets and Robust Bellman Oracles

We give the fixed-policy dual representations for the uncertainty sets defined in Section 3.2. For fixed $\gamma \in (0,1)$ and $v : \mathcal{S} \to \mathbb{R}$, write

$$z_{s,a}(x) := r(s,a,x) + \gamma v(x), \qquad x \in \mathcal{S}.$$

The dependence of $z_{s,a}$ on γ and v is suppressed when it is clear from context.

Kullback-Leibler (KL) Divergence.

Lemma 5 (Lemma C.1 in Chen et al. (2025))**.** *Suppose $\mathcal{P}$ is SA-rectangular and $U = D_{\mathrm{KL}}$ with $\delta > 0$. For any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$, the fixed-policy Bellman operator satisfies*

$$\mathcal{T}^\pi_\gamma(v)(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \sup_{\lambda_a \ge 0} \left\{ -\lambda_a \delta - \lambda_a \log\big(p_{s,a}[e^{-z_{s,a}/\lambda_a}]\big) \right\}.$$

Lemma 6 (Lemma 1 in Li et al. (2026))**.** *Suppose $\mathcal{P}$ is S-rectangular and $U = D_{\mathrm{KL}}$ with $\delta > 0$. The fixed-policy Bellman operator has the following representation for any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$:*

$$\mathcal{T}^\pi_\gamma(v)(s) = \sup_{\lambda>0} \left\{ -\lambda|\mathcal{A}|\delta - \lambda \sum_{a\in\mathcal{A}} \log p_{s,a} \left[\exp\left(-\frac{\pi(a \mid s) z_{s,a}}{\lambda} \right) \right] \right\}.$$

f_k-Divergence.

Lemma 7 (Lemma F.1 in Chen et al. (2025))**.** *Suppose $\mathcal{P}$ is SA-rectangular and $U = D_{f_k}$ with $\delta > 0$. Let $k^* := k/(k-1)$ denote the conjugate exponent, then, for any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$, the fixed-policy Bellman operator satisfies*

$$\mathcal{T}^\pi_\gamma(v)(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \sup_{\theta_a\in\mathbb{R}} \left\{ \theta_a - \big(1 + k(k-1)\delta\big)^{1/k} \big(p_{s,a}[(\theta_a - z_{s,a})_+^{k^*}]\big)^{1/k^*} \right\}.$$

Lemma 8 (Lemma 2 in Li et al. (2026))**.** *Suppose $\mathcal{P}$ is S-rectangular and $U = D_{f_k}$ with $\delta > 0$. Let $k^* := k/(k-1)$ denote the conjugate exponent, then, for any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$, the fixed-policy Bellman operator satisfies*

$$\mathcal{T}^\pi_\gamma(v)(s) = \sup_{\theta\in\mathbb{R}^{|\mathcal{A}|}} \left\{ \sum_{a\in\mathcal{A}} \theta_a - |\mathcal{A}|^{1/k} \big(1 + k(k-1)\delta\big)^{1/k} \left(\sum_{a\in\mathcal{A}} p_{s,a} \left[(\theta_a - \pi(a \mid s) z_{s,a})_+^{k^*} \right] \right)^{1/k^*} \right\}.$$

Total-Variation (TV) distance.

Lemma 9 (Lemma 6 in Iyengar (2005)). *For $\mathcal{P} = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{s,a}(d_{\mathrm{TV}}, \delta)$ with $\delta \in (0,\infty)$, the fixed-policy Bellman operator can be written as follows for any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$:*

$$\mathcal{T}_\gamma^\pi(v)(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \sup_{\mu_a \geq 0} \Big\{ p_{s,a}[z_{s,a} - \mu_a] - \delta \operatorname{Span}(z_{s,a} - \mu_a) \Big\},$$

where each μ_a is a nonnegative vector on $\mathcal{S}$.

Apply the cited L_1-ball duality to each payoff $z_{s,a}$ with L_1 radius 2δ, using $d_{\mathrm{TV}}(q,p) = \frac{1}{2}\|q-p\|_1$, and substitute the resulting dual values into the fixed-policy Bellman operator.

Lemma 10. *Let $\mathcal{P} = \prod_{s\in\mathcal{S}} \mathcal{P}_s(d_{\mathrm{TV}}, \delta)$ with $\delta \in (0,\infty)$. For any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$, the fixed-policy Bellman operator satisfies*

$$\mathcal{T}_\gamma^\pi(v)(s) = \sup_{\mu \in (\mathbb{R}_+^{|\mathcal{S}|})^{\mathcal{A}}} \left\{ \sum_{a\in\mathcal{A}} p_{s,a}[\pi(a \mid s) z_{s,a} - \mu_a] - |\mathcal{A}|\delta \max_{a\in\mathcal{A}} \operatorname{Span}(\pi(a \mid s) z_{s,a} - \mu_a) \right\},$$

where each μ_a is a nonnegative vector on $\mathcal{S}$.

Wasserstein distance.

Lemma 11 (Lemma 5.1 in Wang et al. (2023c)). *Suppose $\mathcal{P} = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{s,a}(W_\ell, \delta)$ with $\delta \in (0,\infty)$. Then, for any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$, the fixed-policy Bellman operator satisfies*

$$\mathcal{T}_\gamma^\pi(v)(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s) \sup_{\lambda_a \geq 0} \left\{ -\lambda_a \delta^\ell + p_{s,a}\left[\inf_{y\in\mathcal{S}} \left(z_{s,a}(y) + \lambda_a \rho(\cdot, y)^\ell \right) \right] \right\}.$$

The cited support-function identity applies to any payoff on $\mathcal{S}$; substitute $z_{s,a} = r(s,a,\cdot) + \gamma v$ for its value vector and combine the action-wise dual values with weights $\pi(a \mid s)$ in the fixed-policy Bellman operator.

Lemma 12. *For $\mathcal{P} = \prod_{s\in\mathcal{S}} \mathcal{P}_s(W_\ell, \delta)$ with $\delta \in (0,\infty)$, the fixed-policy Bellman operator has the following dual form for any $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $v : \mathcal{S} \to \mathbb{R}$:*

$$\mathcal{T}_\gamma^\pi(v)(s) = \sup_{\lambda\geq 0} \left\{ -\lambda|\mathcal{A}|\delta^\ell + \sum_{a\in\mathcal{A}} p_{s,a}\left[\inf_{y\in\mathcal{S}} \left(\pi(a \mid s) z_{s,a}(y) + \lambda \rho(\cdot, y)^\ell \right) \right] \right\}.$$

Proof of Lemmas 5 and 7. Fix $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, $\gamma \in (0,1)$, and $v : \mathcal{S} \to \mathbb{R}$. Although Chen et al. (2025) consider rewards $r(s,a)$, their Lemmas C.1 and F.1 give dual representations for the worst-case expectation of an arbitrary real-valued vector. We apply these identities to $z_{s,a}(x) = r(s,a,x) + \gamma v(x)$, which remains fixed while the transition distribution varies. For $U \in \{D_{\mathrm{KL}}, D_{f_k}\}$,

$$\begin{aligned}
\mathcal{T}_\gamma^\pi(v)(s) &= \sum_{a\in\mathcal{A}} \pi(a \mid s) \inf_{q_{s,a}\in\mathcal{P}_{s,a}(U,\delta)} \sum_{x\in\mathcal{S}} q_{s,a}(x)\big[r(s,a,x) + \gamma v(x)\big] \\
&= \sum_{a\in\mathcal{A}} \pi(a \mid s) \inf_{q_{s,a}\in\mathcal{P}_{s,a}(U,\delta)} \sum_{x\in\mathcal{S}} q_{s,a}(x) z_{s,a}(x).
\end{aligned}$$

Applying their Lemma C.1 to each inner infimum gives the KL representation in Lemma 5. For f_k uncertainty, their Lemma F.1 yields Lemma 7 with the same coefficient $(1 + k(k-1)\delta)^{1/k}$. The dual variables are chosen separately for each action. Transition-dependent rewards thus enter these identities through the payoff vector $z_{s,a}$. In Lemma 5, the expression at $\lambda_a = 0$ is interpreted by its limit as $\lambda_a \downarrow 0$. □

We next prove the two S-rectangular metric dual representations in Lemmas 10 and 12.

Proof of Lemma 10. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$, set $\phi = \pi(\cdot \mid s)$, and write

$$z_{s,a}(x) := r(s,a,x) + \gamma v(x), \qquad a \in \mathcal{A},\ x \in \mathcal{S}.$$

The primal problem is

$$\begin{aligned}\mathcal{T}_\gamma^\pi(v)(s) := \inf_{q_s} & \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)z_{s,a}] \\ \text{s.t.} \quad & q_{s,a} \in \Delta(\mathcal{S}), \quad a \in \mathcal{A}, \\ & \sum_{a\in\mathcal{A}} \|q_{s,a} - p_{s,a}\|_1 \le 2|\mathcal{A}|\delta.\end{aligned} \tag{B.5}$$

Put

$$y_{s,a} := q_{s,a} - p_{s,a}.$$

Since $q_{s,a} = p_{s,a} + y_{s,a}$,

$$q_{s,a} \in \Delta(\mathcal{S}) \iff \begin{cases} y_{s,a} \ge -p_{s,a}, \\ \sum_{x\in\mathcal{S}} y_{s,a}(x) = 0, \end{cases} \qquad \|q_{s,a} - p_{s,a}\|_1 = \|y_{s,a}\|_1.$$

Thus the primal problem in (B.5) becomes

$$\begin{aligned}\mathcal{T}_\gamma^\pi(v)(s) = & \sum_{a\in\mathcal{A}} p_{s,a}[\phi(a)z_{s,a}] \\ & + \min_{y} \sum_{a\in\mathcal{A}}\sum_{x\in\mathcal{S}} y_{s,a}(x)\phi(a)z_{s,a}(x) \\ \text{s.t.} \quad & \sum_{a\in\mathcal{A}} \|y_{s,a}\|_1 \le 2|\mathcal{A}|\delta, \\ & \sum_{x\in\mathcal{S}} y_{s,a}(x) = 0, \quad a \in \mathcal{A}, \\ & y_{s,a} \ge -p_{s,a}, \quad a \in \mathcal{A}.\end{aligned}$$

Introduce $\mu_a \in \mathbb{R}_+^{|\mathcal{S}|}$ and $c_a \in \mathbb{R}$. The dual-norm identity gives

$$\begin{aligned}&\min_{\sum_a \|y_{s,a}\|_1 \le 2|\mathcal{A}|\delta} \sum_{a,x} y_{s,a}(x)\big(\phi(a)z_{s,a}(x) - \mu_a(x) - c_a\big) \\ &\qquad = -2|\mathcal{A}|\delta \max_{a,x} \big|\phi(a)z_{s,a}(x) - \mu_a(x) - c_a\big|.\end{aligned}$$

By the L_1-ball duality of Iyengar (2005, Lemma 6) and finite-dimensional linear-programming strong duality, the primal problem can then be written as

$$\begin{aligned}\mathcal{T}_\gamma^\pi(v)(s) = & \sum_{a\in\mathcal{A}} p_{s,a}[\phi(a)z_{s,a}] \\ & + \sup_{\substack{\mu\in(\mathbb{R}_+^{|\mathcal{S}|})^{\mathcal{A}} \\ c\in\mathbb{R}^{\mathcal{A}}}} \left\{ -\sum_{a\in\mathcal{A}} p_{s,a}[\mu_a] - 2|\mathcal{A}|\delta \max_{a\in\mathcal{A}} \|\phi(a)z_{s,a} - \mu_a - c_a\mathbf{1}\|_\infty \right\}.\end{aligned}$$

$$\inf_{c\in\mathbb{R}^{\mathcal{A}}} \max_{a\in\mathcal{A}} \|\phi(a)z_{s,a} - \mu_a - c_a\mathbf{1}\|_\infty = \frac{1}{2}\max_{a\in\mathcal{A}} \mathrm{Span}(\phi(a)z_{s,a} - \mu_a).$$

Here we use

$$\inf_{c\in\mathbb{R}} \|v - c\mathbf{1}\|_\infty = \frac{1}{2}\,\mathrm{Span}(v).$$

$$\mathcal{T}_\gamma^\pi(v)(s) = \sup_{\mu\in(\mathbb{R}_+^{|\mathcal{S}|})^{\mathcal{A}}} \left\{ \sum_{a\in\mathcal{A}} p_{s,a}[\pi(a \mid s)z_{s,a} - \mu_a] - |\mathcal{A}|\delta \max_{a\in\mathcal{A}} \mathrm{Span}(\pi(a \mid s)z_{s,a} - \mu_a) \right\},$$

which is the claimed dual representation.

□

Proof of Lemma 12. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$, and set $\phi = \pi(\cdot \mid s)$. The primal problem is

$$\begin{aligned}\mathcal{T}_\gamma^\pi(v)(s) = \inf_{q_s} & \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a) z_{s,a}] \\ \text{s.t.} \quad & q_{s,a} \in \Delta(\mathcal{S}), \quad a \in \mathcal{A}, \\ & \sum_{a\in\mathcal{A}} W_\ell(q_{s,a}, p_{s,a})^\ell \le |\mathcal{A}|\delta^\ell.\end{aligned}$$

Let $X := \mathcal{A} \times \mathcal{S}$ and define

$$\bar{p}_s(a,x) := \frac{1}{|\mathcal{A}|} p_{s,a}(x), \qquad \bar{q}_s(a,y) := \frac{1}{|\mathcal{A}|} q_{s,a}(y),$$

and

$$c((a,x),(b,y)) := \begin{cases} \rho(x,y)^\ell, & a = b, \\ +\infty, & a \neq b. \end{cases}$$

$$\inf_{\xi\in\Gamma(\bar{p}_s,\bar{q}_s)} \sum_{a,b\in\mathcal{A}} \sum_{x,y\in\mathcal{S}} c((a,x),(b,y))\xi((a,x),(b,y)) = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} W_\ell(q_{s,a}, p_{s,a})^\ell,$$

and

$$\sum_{a,y} \bar{q}_s(a,y)\phi(a) z_{s,a}(y) = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a) z_{s,a}].$$

Conversely, every finite-cost coupling must preserve the action coordinate. Hence

$$\sum_y \bar{q}_s(a,y) = \sum_x \bar{p}_s(a,x) = \frac{1}{|\mathcal{A}|}, \qquad a \in \mathcal{A},$$

so $q_{s,a}(y) := |\mathcal{A}|\bar{q}_s(a,y)$ belongs to $\Delta(\mathcal{S})$. Thus, the primal problem is $|\mathcal{A}|$ times the corresponding product-space problem with radius δ^ℓ. We adopt the convention $0 \cdot (+\infty) = +\infty$. For each $\lambda \ge 0$, define

$$\varphi_\lambda\big((a,x),(b,y)\big) := -\phi(b) z_{s,b}(y) - \lambda c\big((a,x),(b,y)\big).$$

Since X is finite, the pointwise supremum of φ_λ is measurable. Choosing a maximizing second coordinate for each first coordinate yields a coupling that attains the expectation of this pointwise supremum, while the reverse inequality holds pointwise. Thus φ_λ satisfies the interchangeability principle. By Zhang et al. (2025, Theorem 2.2), applied to the payoff $(a,y) \mapsto -\phi(a) z_{s,a}(y)$,

$$\mathcal{T}_\gamma^\pi(v)(s) = \sup_{\lambda\ge 0} \left\{ -\lambda|\mathcal{A}|\delta^\ell + |\mathcal{A}| \sum_{a,x} \bar{p}_s(a,x) \inf_{b,y} \left(\phi(b) z_{s,b}(y) + \lambda c((a,x),(b,y))\right) \right\}.$$

Since $c((a,x),(b,y)) = +\infty$ for $b \neq a$,

$$\inf_{b,y} \left(\phi(b) z_{s,b}(y) + \lambda c((a,x),(b,y))\right) = \inf_{y\in\mathcal{S}} \left(\phi(a) z_{s,a}(y) + \lambda \rho(x,y)^\ell\right).$$

Substituting $\bar{p}_s(a,x) = |\mathcal{A}|^{-1} p_{s,a}(x)$ gives

$$\begin{aligned}\mathcal{T}_\gamma^\pi(v)(s) = \sup_{\lambda\ge 0} \Bigg\{ & -\lambda|\mathcal{A}|\delta^\ell \\ & + \sum_{a\in\mathcal{A}} p_{s,a} \left[\inf_{y\in\mathcal{S}} \left(\pi(a \mid s) z_{s,a}(y) + \lambda \rho(\cdot, y)^\ell\right) \right] \Bigg\},\end{aligned}$$

which is the claimed dual representation. □

Remark. The proof of Lemma 12 reduces the S-rectangular Wasserstein problem to a standard Wasserstein distributionally robust optimization problem on the enlarged space $\mathcal{A} \times \mathcal{S}$. The key step is to assign infinite transportation cost between points corresponding to different actions. Consequently, any finite-cost coupling is forced to preserve the action coordinate, while transportation

within each action incurs exactly the original state-space cost ρ^ℓ. Under the uniform action marginal $1/|\mathcal{A}|$, the resulting Kantorovich cost becomes

$$K_c(\bar{p}_s, \bar{q}_s) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} W_\ell(q_{s,a}, p_{s,a})^\ell.$$

Thus, the coupled S-rectangular constraint is represented by a single Wasserstein ball on $\mathcal{A} \times \mathcal{S}$. Standard Wasserstein duality can then be applied directly, and the infinite cross-action cost reduces the inner infimum back to an action-wise infimum over $y \in \mathcal{S}$. Thus, the proof converts the coupling across actions in the uncertainty budget into an ordinary Wasserstein dual problem without separating the budget action by action.

C Perturbation Bounds

Based on the dual forms in Appendix B.1, we analyze perturbation bounds for the fixed-policy Bellman operator under eight uncertainty-set and rectangularity combinations. Specifically, for any $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, and $s \in \mathcal{S}$, we aim to bound $\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right|$ in terms of $1 + \mathrm{Span}(v)$, thereby quantifying the effect of replacing the true transition distributions with their empirical counterparts. The resulting bounds are established separately in Corollaries 2–9. Their pointwise estimates use only the relative-error inequalities on $\Omega_{n,\eta}$ and hold for arbitrary v, π, and $\gamma \in (0,1)$; thus the bounds are simultaneous on this event, without a union bound over policies.

Lemma 13. *For any two probability distributions $p, q \in \Delta(\mathcal{S})$ and any bounded vector $h : \mathcal{S} \to \mathbb{R}$,*

$$\left|\sum_{x \in \mathcal{S}} (p(x) - q(x))h(x)\right| \le d_{\mathrm{TV}}(p,q)\,\mathrm{Span}(h).$$

Proof. Let $M = \max_{x \in \mathcal{S}} h(x)$ and $m = \min_{x \in \mathcal{S}} h(x)$, and set $\bar{h}(x) = h(x) - (M+m)/2$. Then $\sum_x (p(x) - q(x))\bar{h}(x) = \sum_x (p(x) - q(x))h(x)$, and

$$\max_{x \in \mathcal{S}} |\bar{h}(x)| \;=\; \frac{M-m}{2} = \frac{\mathrm{Span}(h)}{2}.$$

Hence

$$\left|\sum_x (p-q)(x)h(x)\right| \le \sum_x |p(x) - q(x)|\,|\bar{h}(x)| \le \|p-q\|_1 \frac{\mathrm{Span}(h)}{2} = d_{\mathrm{TV}}(p,q)\,\mathrm{Span}(h).$$

The last equality uses $d_{\mathrm{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$. □

We will use this inequality repeatedly below.

Corollary 2. *Under the SA-rectangular KL uncertainty set, on the good event $\Omega_{n,\eta}$ with $\eta \le 1/2$, simultaneously for every $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $\gamma \in (0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le 3\eta\big(1 + \mathrm{Span}(v)\big).$$

Proof of Corollary 2. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$. The fixed-policy operator definition and its empirical counterpart give

$$\mathcal{T}_\gamma^\pi(v)(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \inf_{q_{s,a} \in \mathcal{P}_{s,a}(D_{\mathrm{KL}}, \delta)} q_{s,a}[r(s,a,\cdot) + \gamma v],$$

$$\widehat{\mathcal{T}}_\gamma^\pi(v)(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \inf_{q_{s,a} \in \widehat{\mathcal{P}}_{s,a}(D_{\mathrm{KL}}, \delta)} q_{s,a}[r(s,a,\cdot) + \gamma v].$$

Using the dual representation in Lemma 5, on $\Omega_{n,\eta}$, for each $a \in \mathcal{A}$, Chen et al.'s Lemma C.1 provides the KL primal–dual equivalence, while their Lemma C.6 bounds the change of the resulting dual objective when $p_{s,a}$ is replaced by $\widehat{p}_{s,a}$; see Chen et al. (2025, Lemmas C.1 and C.6).

$$\left|\inf_{q_{s,a} \in \widehat{\mathcal{P}}_{s,a}(D_{\mathrm{KL}}, \delta)} q_{s,a}[r(s,a,\cdot) + \gamma v] - \inf_{q_{s,a} \in \mathcal{P}_{s,a}(D_{\mathrm{KL}}, \delta)} q_{s,a}[r(s,a,\cdot) + \gamma v]\right|$$
$$\le 3\eta\ \mathrm{Span}(r(s,a,\cdot) + \gamma v).$$

Weighting these action-wise estimates by $\pi(a \mid s)$ gives

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\le \sum_{a\in\mathcal{A}} \pi(a \mid s) \left| \inf_{q_{s,a}\in\widehat{\mathcal{P}}_{s,a}(D_{\mathrm{KL}},\delta)} q_{s,a}[r(s,a,\cdot)+\gamma v] - \inf_{q_{s,a}\in\mathcal{P}_{s,a}(D_{\mathrm{KL}},\delta)} q_{s,a}[r(s,a,\cdot)+\gamma v] \right| \\
&\le 3\eta \sum_{a\in\mathcal{A}} \pi(a \mid s)\, \mathrm{Span}(r(s,a,\cdot)+\gamma v) \\
&\le 3\eta\big(1+\mathrm{Span}(v)\big).
\end{aligned}$$

The last inequality uses $\sum_a \pi(a \mid s) = 1$. □

Corollary 3. *Under the S-rectangular KL uncertainty set, on the good event $\Omega_{n,\eta}$ with $\eta \le 1/2$, simultaneously for every $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $\gamma \in (0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le 3\eta\big(1+\mathrm{Span}(v)\big).$$

Proof of Corollary 3. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$, and set $\phi = \pi(\cdot \mid s)$. Using the dual representation in Lemma 6, the common term $-\lambda|\mathcal{A}|\delta$ cancels, and hence

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\le \sup_{\lambda>0} \left| \sum_{a\in\mathcal{A}} \lambda \log \widehat{p}_{s,a}\left[\exp\left(-\frac{\phi(a)(r(s,a,\cdot)+\gamma v)}{\lambda}\right)\right] \right. \\
&\qquad\qquad \left. - \sum_{a\in\mathcal{A}} \lambda \log p_{s,a}\left[\exp\left(-\frac{\phi(a)(r(s,a,\cdot)+\gamma v)}{\lambda}\right)\right] \right| \\
&\le \sum_{a\in\mathcal{A}} \sup_{\lambda>0} \left| \lambda \log \widehat{p}_{s,a}\left[\exp\left(-\frac{\phi(a)(r(s,a,\cdot)+\gamma v)}{\lambda}\right)\right] \right. \\
&\qquad\qquad \left. - \lambda \log p_{s,a}\left[\exp\left(-\frac{\phi(a)(r(s,a,\cdot)+\gamma v)}{\lambda}\right)\right] \right|.
\end{aligned}$$

On $\Omega_{n,\eta}$, Lemma C.6 of Chen et al. (2025) gives, for every $a \in \mathcal{A}$,

$$\begin{aligned}
&\sup_{\lambda>0} \left| \lambda \log \widehat{p}_{s,a}\left[\exp\left(-\frac{\phi(a)(r(s,a,\cdot)+\gamma v)}{\lambda}\right)\right] \right. \\
&\qquad\qquad \left. - \lambda \log p_{s,a}\left[\exp\left(-\frac{\phi(a)(r(s,a,\cdot)+\gamma v)}{\lambda}\right)\right] \right| \\
&\le 3\eta\ \mathrm{Span}\big(\phi(a)(r(s,a,\cdot)+\gamma v)\big) \\
&= 3\eta\,\phi(a)\,\mathrm{Span}\big(r(s,a,\cdot)+\gamma v\big).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| &\le 3\eta \sum_{a\in\mathcal{A}} \phi(a)\,\mathrm{Span}\big(r(s,a,\cdot)+\gamma v\big) \\
&\le 3\eta \max_{a\in\mathcal{A}} \mathrm{Span}\big(r(s,a,\cdot)+\gamma v\big) \\
&\le 3\eta\big(1+\mathrm{Span}(v)\big).
\end{aligned}$$

This proves the fixed-policy bound with the same coupled uncertainty budget. □

We next develop a common perturbation argument for the f_k-divergence uncertainty sets. The following lemmas describe the coupled dual problem for an arbitrary finite action set. We then apply them to each action separately in the SA-rectangular case and to the full action set in the S-rectangular case.

To this end, we define $p_{t,s,a} = t\widehat{p}_{s,a} + (1-t)p_{s,a}$ and study the corresponding dual value $G(t)$, where $G(0)$ and $G(1)$ are the true and empirical dual values, respectively. Lemma 14 shows that the supremum defining $G(t)$ can be restricted to the same compact set X for all $t \in [0,1]$. Lemma 15 establishes the existence of $X^*(t)$, the regularity of $f(\theta,t)$, and a uniform bound on $f_t(\theta,t)$, while Lemma 16 expresses $G(1) - G(0)$ as the integral of $f_t(\theta^*(t),t)$. Lemma 17 gives the identity satisfied by each optimizer $\theta^*(t)$, which is then used in Lemma 18 to show that

$$\left|f_t\big(\theta^*(t),t\big)\right| \le 2\eta \max_{a\in\mathcal{A}} \mathrm{Span}(z_{s,a}).$$

Combining these results yields the perturbation bounds in Corollaries 4 and 5.

Lemma 14. *Fix $s \in \mathcal{S}$, $\phi \in \Delta(\mathcal{A})$, $k > 1$, $k^* = k/(k-1)$, and $\delta > 0$. Let*

$$c := |\mathcal{A}|^{1/k}\big(1 + k(k-1)\delta\big)^{1/k}, \quad p_{t,s,a} := t\widehat{p}_{s,a} + (1-t)p_{s,a}, \quad \Delta_{s,a} := \widehat{p}_{s,a} - p_{s,a}.$$

For $\theta \in \mathbb{R}^{|\mathcal{A}|}$ and $t \in [0,1]$, define

$$f(\theta,t) := \sum_{a\in\mathcal{A}} \theta_a - c\left(\sum_{a\in\mathcal{A}} p_{t,s,a}\left[(\theta_a - \phi(a)z_{s,a})_+^{k^*}\right]\right)^{1/k^*},$$
$$G(t) := \sup_{\theta\in\mathbb{R}^{|\mathcal{A}|}} f(\theta,t).$$

For every $a \in \mathcal{A}$, let

$$\mathcal{X}_{s,a} := \mathrm{supp}(\widehat{p}_{s,a}) \cup \mathrm{supp}(p_{s,a}),$$
$$L_a := \min_{x\in\mathcal{X}_{s,a}} \phi(a)z_{s,a}(x), \qquad M_a := \max_{x\in\mathcal{X}_{s,a}} \phi(a)z_{s,a}(x),$$

and define

$$X := \prod_{a\in\mathcal{A}} \left[L_a,\ M_a + \frac{\sum_{a\in\mathcal{A}}(M_a - L_a)}{c - |\mathcal{A}|^{1/k}}\right].$$

Then, for every $t \in [0,1]$,

$$G(t) = \sup_{\theta\in X} f(\theta,t).$$

Proof. Fix $t \in [0,1]$. If $\theta_a \le L_a$, then, for every $x \in \mathcal{X}_{s,a}$,

$$\big(\theta_a - \phi(a)z_{s,a}(x)\big)_+ = \big(L_a - \phi(a)z_{s,a}(x)\big)_+ = 0.$$

Thus, replacing θ_a by L_a leaves the second term of $f(\theta,t)$ unchanged and weakly increases its first term. Hence it suffices to consider $\theta_a \ge L_a$ for all $a \in \mathcal{A}$.

Suppose that, for some $a_0 \in \mathcal{A}$,

$$\theta_{a_0} > M_{a_0} + \frac{\sum_{a\in\mathcal{A}}(M_a - L_a)}{c - |\mathcal{A}|^{1/k}}.$$

For every $a \in \mathcal{A}$ and $x \in \mathcal{X}_{s,a}$,

$$\big(\theta_a - \phi(a)z_{s,a}(x)\big)_+ \ge (\theta_a - M_a)_+.$$

Consequently,

$$f(\theta,t) \le \sum_{a\in\mathcal{A}} M_a + \sum_{a\in\mathcal{A}} (\theta_a - M_a)_+ - c\left(\sum_{a\in\mathcal{A}} (\theta_a - M_a)_+^{k^*}\right)^{1/k^*}.$$

By Hölder's inequality,

$$\sum_{a\in\mathcal{A}} (\theta_a - M_a)_+ \le |\mathcal{A}|^{1/k}\left(\sum_{a\in\mathcal{A}} (\theta_a - M_a)_+^{k^*}\right)^{1/k^*}.$$

Therefore,

$$\begin{aligned} f(\theta,t) &\le \sum_{a\in\mathcal{A}} M_a - \big(c-|\mathcal{A}|^{1/k}\big)\left(\sum_{a\in\mathcal{A}}(\theta_a-M_a)_+^{k^*}\right)^{1/k^*} \\ &\le \sum_{a\in\mathcal{A}} M_a - \big(c-|\mathcal{A}|^{1/k}\big)(\theta_{a_0}-M_{a_0}) \\ &< \sum_{a\in\mathcal{A}} L_a = f\big((L_a)_{a\in\mathcal{A}},t\big). \end{aligned}$$

Thus no point outside X can attain the supremum, which proves the claim. □

Lemma 15. *Under the notation of Lemma 14, assume that, for every $a\in\mathcal{A}$ and $t\in(0,1)$,*

$$p_{t,s,a} \ge \frac{1}{2}p_{s,a}, \qquad \left\|\frac{\Delta_{s,a}}{p_{t,s,a}}\right\|_{L^\infty(p_{t,s,a})} \le 2\eta.$$

For $\theta\in X$, define

$$B_\theta(t) := \sum_{a\in\mathcal{A}} p_{t,s,a}\left[(\theta_a-\phi(a)z_{s,a})_+^{k^*}\right], \qquad X^*(t) := \arg\max_{\theta\in X} f(\theta,t).$$

Then the following hold.

1. *The set X is compact and*
$$X^*(t)\neq\varnothing, \qquad \forall t\in[0,1].$$
2. *For every $\theta\in X$, $f(\theta,\cdot)$ is absolutely continuous on $[0,1]$, differentiable on $(0,1)$, and*
$$|f_t(\theta,t)| = \begin{cases} 0, & B_\theta(0)=B_\theta(1)=0, \\ \dfrac{c}{k^*}\left(tB_\theta(1)+(1-t)B_\theta(0)\right)^{\frac{1}{k^*}-1}\left|B_\theta(1)-B_\theta(0)\right|, & \text{otherwise}. \end{cases}$$
3. *There exists a finite constant B such that*
$$|f_t(\theta,t)|\le B, \qquad \forall t\in(0,1), \quad \forall\theta\in X.$$

Proof. The set X is a finite product of closed bounded intervals. For every $t\in[0,1]$, $f(\cdot,t)$ is continuous on X; hence X is compact and Weierstrass' theorem gives $X^*(t)\neq\varnothing$.

Fix $\theta\in X$. Since $p_{t,s,a}=t\widehat{p}_{s,a}+(1-t)p_{s,a}$,

$$B_\theta(t) = tB_\theta(1)+(1-t)B_\theta(0).$$

Because $u\mapsto u^{1/k^*}$ is concave on $\mathbb{R}_+$, $f(\theta,\cdot)$ is finite and convex on $[0,1]$, and is therefore absolutely continuous. If $B_\theta(t_0)=0$ for some $t_0\in(0,1)$, then

$$B_\theta(0)=B_\theta(1)=0, \qquad B_\theta(t)=0, \quad \forall t\in[0,1].$$

Thus $f_t(\theta,t)=0$. Otherwise, $B_\theta(t)>0$ on $(0,1)$, and

$$|f_t(\theta,t)| = \frac{c}{k^*}\left(tB_\theta(1)+(1-t)B_\theta(0)\right)^{\frac{1}{k^*}-1}\left|B_\theta(1)-B_\theta(0)\right|.$$

Moreover,

$$\begin{aligned} |B_\theta(1)-B_\theta(0)| &= \left|\sum_{a\in\mathcal{A}}\Delta_{s,a}\left[(\theta_a-\phi(a)z_{s,a})_+^{k^*}\right]\right| \\ &\le 2\eta\sum_{a\in\mathcal{A}} p_{t,s,a}\left[(\theta_a-\phi(a)z_{s,a})_+^{k^*}\right] \\ &= 2\eta B_\theta(t). \end{aligned}$$

For every $\theta \in X$ and $t \in [0,1]$,

$$B_\theta(t) \le \sum_{a\in\mathcal{A}} \left(M_a - L_a + \frac{\sum_{b\in\mathcal{A}}(M_b - L_b)}{c - |\mathcal{A}|^{1/k}} \right)^{k^*}.$$

Hence

$$|f_t(\theta,t)| \le \frac{2\eta c}{k^*} \left(\sum_{a\in\mathcal{A}} \left(M_a - L_a + \frac{\sum_{b\in\mathcal{A}}(M_b - L_b)}{c - |\mathcal{A}|^{1/k}} \right)^{k^*} \right)^{1/k^*},$$

which is finite and uniform over $t \in (0,1)$ and $\theta \in X$. □

Lemma 16. *Under the conditions of Lemma 15, for any measurable selection $\theta^*(t) \in X^*(t)$,*

$$G(1) = G(0) + \int_0^1 f_t\big(\theta^*(r), r\big)\, dr.$$

Proof. By Lemma 15,

$$X^*(t) \neq \varnothing, \qquad \forall t \in [0,1],$$

and, for every $\theta \in X$, $f(\theta,\cdot)$ is absolutely continuous on $[0,1]$, differentiable on $(0,1)$, and satisfies

$$|f_t(\theta,t)| \le B, \qquad \forall t \in (0,1), \quad \forall \theta \in X.$$

Since X is compact and f is jointly continuous, $X^*(t)$ admits a measurable selection. Thus the hypotheses of Milgrom & Segal (2002, Theorem 2) hold with the integrable envelope $t \mapsto B$, and hence

$$G(1) = G(0) + \int_0^1 f_t\big(\theta^*(r), r\big)\, dr.$$

□

Lemma 17. *Let $\theta^*(t) \in X^*(t)$ and suppose that $B_{\theta^*(t)}(t) > 0$. Then, for every $a \in \mathcal{A}$,*

$$1 = cB_{\theta^*(t)}(t)^{\frac{1}{k^*}-1} p_{t,s,a} \left[(\theta^*_a(t) - \phi(a) z_{s,a})_+^{k^*-1} \right].$$

Proof. By Lemma 14, $\theta^*(t)$ maximizes $f(\cdot,t)$ over $\mathbb{R}^{|\mathcal{A}|}$. Since $k^* > 1$ and $B_{\theta^*(t)}(t) > 0$,

$$\begin{aligned} 0 &= f_{\theta_a}(\theta^*(t), t) \\ &= 1 - cB_{\theta^*(t)}(t)^{\frac{1}{k^*}-1} p_{t,s,a} \left[(\theta^*_a(t) - \phi(a) z_{s,a})_+^{k^*-1} \right]. \end{aligned}$$

Rearranging proves the result. □

Lemma 18. *Under the conditions of Lemma 15, for every $\theta^*(t) \in X^*(t)$ and every $t \in (0,1)$,*

$$\big|f_t\big(\theta^*(t), t\big)\big| \le 2\eta \max_{a\in\mathcal{A}} \mathrm{Span}(z_{s,a}) \le 2\eta\big(1 + \mathrm{Span}(v)\big).$$

Proof. If $\theta^*(t) = (L_a)_{a\in\mathcal{A}}$, then $B_{\theta^*(t)}(r) = 0$ for every $r \in [0,1]$, and hence

$$f_t\big(\theta^*(t), t\big) = 0.$$

Otherwise, $B_{\theta^*(t)}(t) > 0$. Indeed, if it were zero, then, since $t \in (0,1)$ and $\mathrm{supp}(p_{t,s,a}) = \mathcal{X}_{s,a}$,

$$\theta^*_a(t) \le \phi(a) z_{s,a}(x), \qquad \forall a \in \mathcal{A}, \quad \forall x \in \mathcal{X}_{s,a}.$$

Thus $\theta^*_a(t) \le L_a$ for every a, which, together with $\theta^*(t) \in X$, implies $\theta^*(t) = (L_a)_{a\in\mathcal{A}}$.

Write $\theta^* = \theta^*(t)$. Since $\Delta_{s,a}[\mathbf{1}] = 0$,

$$\begin{aligned} |B_{\theta^*}(1) - B_{\theta^*}(0)| &= \left| \sum_{a\in\mathcal{A}} \Delta_{s,a} \left[(\theta^*_a - \phi(a) z_{s,a})_+^{k^*} - (\theta^*_a - M_a)_+^{k^*} \right] \right| \\ &\le 2\eta \sum_{a\in\mathcal{A}} p_{t,s,a} \left[\left| (\theta^*_a - \phi(a) z_{s,a})_+^{k^*} - (\theta^*_a - M_a)_+^{k^*} \right| \right]. \end{aligned}$$

For every $x \in \mathcal{X}_{s,a}$,

$$\begin{aligned}
&\left|(\theta_a^* - \phi(a) z_{s,a}(x))_+^{k^*} - (\theta_a^* - M_a)_+^{k^*}\right| \\
&\quad \le k^* \, (\theta_a^* - \phi(a) z_{s,a}(x))_+^{k^*-1} \operatorname{Span}\left((\theta_a^* - \phi(a) z_{s,a})_+\right) \\
&\quad \le k^* \, (\theta_a^* - \phi(a) z_{s,a}(x))_+^{k^*-1} \, \phi(a) \operatorname{Span}(z_{s,a}),
\end{aligned}$$

where the first inequality is the mean-value theorem. Therefore,

$$\begin{aligned}
\left|f_t\big(\theta^*(t), t\big)\right| &\le 2\eta c B_{\theta^*}(t)^{\frac{1}{k^*}-1} \sum_{a \in \mathcal{A}} p_{t,s,a} \left[(\theta_a^* - \phi(a) z_{s,a})_+^{k^*-1}\right] \phi(a) \operatorname{Span}(z_{s,a}) \\
&= 2\eta \sum_{a \in \mathcal{A}} \phi(a) \operatorname{Span}(z_{s,a}) \\
&\le 2\eta \max_{a \in \mathcal{A}} \operatorname{Span}(z_{s,a}) \\
&\le 2\eta\big(1 + \operatorname{Span}(v)\big),
\end{aligned}$$

where the equality uses Lemma 17. □

We first apply these lemmas to the separate local problems in the SA-rectangular model.

Corollary 4. *Under the SA-rectangular f_k-divergence uncertainty set, on the good event $\Omega_{n,\eta}$ with $\eta \le 1/2$, simultaneously for every $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $\gamma \in (0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le 2\eta\big(1 + \operatorname{Span}(v)\big).$$

Proof of Corollary 4. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$, and first suppose $\delta > 0$. For each fixed $a \in \mathcal{A}$, apply Lemmas 14–18 to an auxiliary problem whose action set is the singleton $\{a\}$. Its only action distribution has $\phi(a) = 1$, and its dual coefficient is $c = (1 + k(k-1)\delta)^{1/k} > 1$. Thus its dual value is

$$G_a(t) := \sup_{\theta_a \in \mathbb{R}} \left\{ \theta_a - c \left(p_{t,s,a} \left[(\theta_a - z_{s,a})_+^{k^*} \right] \right)^{1/k^*} \right\}.$$

This is exactly the local dual problem in Lemma 7, with center $p_{t,s,a}$. The specialization uses a singleton action set, rather than an action distribution with $\phi(a) = 1$ in the original multi-action S-rectangular problem.

Since the samples are drawn from $p_{s,a}$, $\operatorname{supp}(\widehat{p}_{s,a}) \subseteq \operatorname{supp}(p_{s,a})$ almost surely. On $\Omega_{n,\eta}$, the supports coincide, and for every $t \in (0,1)$,

$$\begin{aligned}
p_{t,s,a} &= p_{s,a} + t\Delta_{s,a} \ge (1-\eta) p_{s,a} \ge \tfrac{1}{2} p_{s,a}, \\
\left\| \frac{\Delta_{s,a}}{p_{t,s,a}} \right\|_{L^\infty(p_{t,s,a})} &\le \frac{1}{1-\eta} \left\| \frac{\Delta_{s,a}}{p_{s,a}} \right\|_{L^\infty(p_{s,a})} \le 2\eta.
\end{aligned}$$

Hence the conditions of Lemma 15 hold. Lemmas 16–18 then give

$$|G_a(1) - G_a(0)| \le 2\eta \operatorname{Span}(z_{s,a}).$$

By Lemma 7 and its empirical counterpart, $\mathcal{T}_\gamma^\pi(v)(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) G_a(0)$ and $\widehat{\mathcal{T}}_\gamma^\pi(v)(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) G_a(1)$. Since $\sum_a \pi(a \mid s) = 1$, we obtain

$$\begin{aligned}
\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| &\le \sum_{a \in \mathcal{A}} \pi(a \mid s) |G_a(1) - G_a(0)| \\
&\le 2\eta \sum_{a \in \mathcal{A}} \pi(a \mid s) \operatorname{Span}(z_{s,a}) \\
&\le 2\eta\big(1 + \operatorname{Span}(v)\big).
\end{aligned}$$

If $\delta = 0$, each local uncertainty set reduces to its center. On $\Omega_{n,\eta}$, $\|\widehat{p}_{s,a} - p_{s,a}\|_1 \le \eta$, so Lemma 13 gives $|\widehat{p}_{s,a}[z_{s,a}] - p_{s,a}[z_{s,a}]| \le (\eta/2) \operatorname{Span}(z_{s,a})$. Weighting by $\pi(a \mid s)$ and summing over actions again yields the fixed-policy bound. □

We now apply the same lemmas to the full action set in the S-rectangular model, retaining the coupled dual problem for $\phi = \pi(\cdot \mid s)$.

Corollary 5. *Under the S-rectangular f_k-divergence uncertainty set, on the good event $\Omega_{n,\eta}$ with $\eta \le 1/2$, simultaneously for every $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $\gamma \in (0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le 2\eta\big(1 + \mathrm{Span}(v)\big).$$

Proof of Corollary 5. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$, and set $\phi = \pi(\cdot \mid s)$. The case $\delta = 0$ follows from the same relative-error argument used in Corollary 4, weighting the action-wise estimates by $\pi(a \mid s)$ and summing, so suppose $\delta > 0$. Let G_ϕ denote the function G in Lemma 14. By Lemma 8 and its empirical counterpart,

$$\mathcal{T}_\gamma^\pi(v)(s) = G_\phi(0), \qquad \widehat{\mathcal{T}}_\gamma^\pi(v)(s) = G_\phi(1).$$

On $\Omega_{n,\eta}$, for every $a \in \mathcal{A}$ and $t \in (0,1)$,

$$p_{t,s,a} = p_{s,a} + t\Delta_{s,a} \ge (1-\eta)p_{s,a} \ge \frac{1}{2}p_{s,a},$$

$$\left\|\frac{\Delta_{s,a}}{p_{t,s,a}}\right\|_{L^\infty(p_{t,s,a})} \le \frac{1}{1-\eta}\left\|\frac{\Delta_{s,a}}{p_{s,a}}\right\|_{L^\infty(p_{s,a})} \le 2\eta.$$

Thus Lemma 15 applies. For a measurable selection $\theta^*(t) \in X^*(t)$, Lemmas 16 and 18 give

$$\begin{aligned}|G_\phi(1) - G_\phi(0)| &\le \int_0^1 \left|f_t\big(\theta^*(r), r\big)\right| dr \\ &\le 2\eta\big(1 + \mathrm{Span}(v)\big).\end{aligned}$$

This gives the fixed-policy bound. □

Corollary 6. *Under the SA-rectangular TV uncertainty set, on the good event $\Omega_{n,\eta}$, simultaneously for every $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $\gamma \in (0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le \frac{\eta}{2}\big(1 + \mathrm{Span}(v)\big).$$

Proof of Corollary 6. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$, and first suppose $\delta > 0$. By Lemma 9, let $\widehat{\mu}_a, \mu_a \ge 0$ be optimizers for the empirical and true dual objectives for $z_{s,a}$, respectively. Then

$$\begin{aligned}\widehat{p}_{s,a}[z_{s,a} - \widehat{\mu}_a] - \delta\,\mathrm{Span}(z_{s,a} - \widehat{\mu}_a) &\ge \widehat{p}_{s,a}[z_{s,a}] - \delta\,\mathrm{Span}(z_{s,a}), \\ \delta(\mathrm{Span}(z_{s,a}) - \mathrm{Span}(z_{s,a} - \widehat{\mu}_a)) &\ge \widehat{p}_{s,a}[\widehat{\mu}_a] \ge 0, \\ p_{s,a}[z_{s,a} - \mu_a] - \delta\,\mathrm{Span}(z_{s,a} - \mu_a) &\ge p_{s,a}[z_{s,a}] - \delta\,\mathrm{Span}(z_{s,a}), \\ \delta(\mathrm{Span}(z_{s,a}) - \mathrm{Span}(z_{s,a} - \mu_a)) &\ge p_{s,a}[\mu_a] \ge 0.\end{aligned}$$

Moreover, on $\Omega_{n,\eta}$,

$$d_{\mathrm{TV}}\big(\widehat{p}_{s,a}, p_{s,a}\big) = \frac{1}{2}\left\|\widehat{p}_{s,a} - p_{s,a}\right\|_1 \le \frac{\eta}{2}.$$

Therefore, by Lemma 13,

$$\begin{aligned}&\left|\inf_{q_{s,a} \in \widehat{\mathcal{P}}_{s,a}(d_{\mathrm{TV}},\delta)} q_{s,a}[z_{s,a}] - \inf_{q_{s,a} \in \mathcal{P}_{s,a}(d_{\mathrm{TV}},\delta)} q_{s,a}[z_{s,a}]\right| \\ &\quad\le \max\left\{\left|(\widehat{p}_{s,a} - p_{s,a})[z_{s,a} - \widehat{\mu}_a]\right|, \left|(\widehat{p}_{s,a} - p_{s,a})[z_{s,a} - \mu_a]\right|\right\} \\ &\quad\le d_{\mathrm{TV}}\big(\widehat{p}_{s,a}, p_{s,a}\big)\max\{\mathrm{Span}(z_{s,a} - \widehat{\mu}_a), \mathrm{Span}(z_{s,a} - \mu_a)\} \\ &\quad\le d_{\mathrm{TV}}\big(\widehat{p}_{s,a}, p_{s,a}\big)\,\mathrm{Span}(z_{s,a}) \le \frac{\eta}{2}\,\mathrm{Span}(z_{s,a}).\end{aligned}$$

If $\delta = 0$, each uncertainty set reduces to its center, and the same action-wise bound follows directly from Lemma 13. Weighting these estimates by $\pi(a \mid s)$ and using $\sum_a \pi(a \mid s) = 1$ yields

$$\begin{aligned}\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| &\le \frac{\eta}{2}\sum_{a \in \mathcal{A}} \pi(a \mid s)\,\mathrm{Span}\big(r(s,a,\cdot) + \gamma v\big) \\ &\le \frac{\eta}{2}\big(1 + \mathrm{Span}(v)\big).\end{aligned}$$

□

Corollary 7. *Under the S-rectangular TV uncertainty set, on the good event $\Omega_{n,\eta}$, simultaneously for every $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $\gamma \in (0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le \frac{\eta}{2}\big(1 + \mathrm{Span}(v)\big).$$

Proof of Corollary 7. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$, and set $\phi = \pi(\cdot \mid s)$. By Lemma 10, let $\widehat{\mu}_a, \mu_a \ge 0$ be optimizers for the empirical and true dual objectives, respectively. Replace μ_a by

$$\min\left\{\mu_a - \min_{y\in\mathcal{S}} \mu_a(y),\ \phi(a)z_{s,a} - \min_{y\in\mathcal{S}} \phi(a)z_{s,a}(y)\right\}.$$

Then

$$\phi(a)z_{s,a} - \mu_a = \max\left\{\phi(a)z_{s,a} - \mu_a + \min_{y\in\mathcal{S}} \mu_a(y),\ \min_{y\in\mathcal{S}} \phi(a)z_{s,a}(y)\right\}.$$

The resulting vector is pointwise no smaller than the original one, and its values lie between

$$\min_{y\in\mathcal{S}} \phi(a)z_{s,a}(y) \quad \text{and} \quad \max_{y\in\mathcal{S}} \phi(a)z_{s,a}(y).$$

Thus its expectation does not decrease and its span does not increase. Consequently, an optimizer can be chosen such that

$$\mathrm{Span}(\phi(a)z_{s,a} - \mu_a) \le \phi(a)\,\mathrm{Span}(z_{s,a}).$$

The same argument applies to $\widehat{\mu}_a$:

$$\mathrm{Span}(\phi(a)z_{s,a} - \widehat{\mu}_a) \le \phi(a)\,\mathrm{Span}(z_{s,a}).$$

On $\Omega_{n,\eta}$, Lemma 13 gives

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\quad\le \sum_{a\in\mathcal{A}} d_{\mathrm{TV}}\big(\widehat{p}_{s,a}, p_{s,a}\big)\phi(a)\,\mathrm{Span}(z_{s,a}) \\
&\quad\le \frac{\eta}{2}\sum_{a\in\mathcal{A}} \phi(a)\,\mathrm{Span}(z_{s,a}) \\
&\quad\le \frac{\eta}{2}\max_{a\in\mathcal{A}} \mathrm{Span}(z_{s,a}).
\end{aligned}$$

Since $\phi = \pi(\cdot \mid s)$, this gives

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\quad\le \frac{\eta}{2}\max_{a\in\mathcal{A}} \mathrm{Span}\big(r(s,a,\cdot) + \gamma v\big) \\
&\quad\le \frac{\eta}{2}\big(1 + \mathrm{Span}(v)\big).
\end{aligned}$$

□

Corollary 8. *Under the SA-rectangular Wasserstein uncertainty set, on the good event $\Omega_{n,\eta}$, simultaneously for every $v \in \mathbb{R}^{|\mathcal{S}|}$, $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$, and $\gamma \in (0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le \frac{\eta}{2}\big(1 + \mathrm{Span}(v)\big).$$

Proof of Corollary 8. Fix $\pi \in \Pi_{\mathrm{SR}}$ and $s \in \mathcal{S}$. By Lemma 11 and its empirical counterpart, weighting the action-wise dual estimates by $\pi(a \mid s)$ gives

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\quad\le \sum_{a\in\mathcal{A}} \pi(a \mid s)\ \sup_{\lambda_a \ge 0}\left|\big(\widehat{p}_{s,a} - p_{s,a}\big)\left[\inf_{y\in\mathcal{S}}\big(z_{s,a}(y) + \lambda_a \rho(\cdot, y)^\ell\big)\right]\right|.
\end{aligned}$$

Since $\rho(x,x)=0$, for every $x\in\mathcal{S}$ and $\lambda_a\geq 0$,

$$\min_{y\in\mathcal{S}} z_{s,a}(y) \leq \inf_{y\in\mathcal{S}} \left(z_{s,a}(y)+\lambda_a\rho(x,y)^\ell\right) \leq z_{s,a}(x) \leq \max_{y\in\mathcal{S}} z_{s,a}(y).$$

Hence, by Lemma 13 and $d_{\mathrm{TV}}(\widehat{p}_{s,a},p_{s,a})\leq\eta/2$ on $\Omega_{n,\eta}$, together with $\sum_a \pi(a\mid s)=1$,

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s)-\mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\quad\leq \frac{\eta}{2}\sum_{a\in\mathcal{A}}\pi(a\mid s)\,\mathrm{Span}(z_{s,a}) \\
&\quad\leq \frac{\eta}{2}\sum_{a\in\mathcal{A}}\pi(a\mid s)\,\mathrm{Span}\big(r(s,a,\cdot)+\gamma v\big) \\
&\quad\leq \frac{\eta}{2}\big(1+\mathrm{Span}(v)\big).
\end{aligned}$$

□

Corollary 9. *Under the S-rectangular Wasserstein uncertainty set, on the good event $\Omega_{n,\eta}$, simultaneously for every $v\in\mathbb{R}^{|\mathcal{S}|}$, $\pi\in\Pi_{\mathrm{SR}}$, $s\in\mathcal{S}$, and $\gamma\in(0,1)$,*

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s)-\mathcal{T}_\gamma^\pi(v)(s)\right| \leq \frac{\eta}{2}\big(1+\mathrm{Span}(v)\big).$$

Proof of Corollary 9. Fix $\pi\in\Pi_{\mathrm{SR}}$ and $s\in\mathcal{S}$, and set $\phi=\pi(\cdot\mid s)$. By Lemma 12 and its empirical counterpart,

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s)-\mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\quad\leq \sup_{\lambda\geq 0}\left|\sum_{a\in\mathcal{A}}\left(\widehat{p}_{s,a}-p_{s,a}\right)\left[\inf_{y\in\mathcal{S}}\left(\phi(a)z_{s,a}(y)+\lambda\rho(\cdot,y)^\ell\right)\right]\right| \\
&\quad\leq \sup_{\lambda\geq 0}\sum_{a\in\mathcal{A}} d_{\mathrm{TV}}\left(\widehat{p}_{s,a},p_{s,a}\right)\mathrm{Span}\left(\inf_{y\in\mathcal{S}}\left(\phi(a)z_{s,a}(y)+\lambda\rho(\cdot,y)^\ell\right)\right).
\end{aligned}$$

Since $\rho(x,x)=0$, for every $x\in\mathcal{S}$,

$$\min_{y\in\mathcal{S}}\phi(a)z_{s,a}(y) \leq \inf_{y\in\mathcal{S}}\left(\phi(a)z_{s,a}(y)+\lambda\rho(x,y)^\ell\right) \leq \phi(a)z_{s,a}(x).$$

Hence, on $\Omega_{n,\eta}$,

$$\begin{aligned}
&\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s)-\mathcal{T}_\gamma^\pi(v)(s)\right| \\
&\quad\leq \frac{\eta}{2}\sum_{a\in\mathcal{A}}\phi(a)\,\mathrm{Span}(z_{s,a}) \\
&\quad\leq \frac{\eta}{2}\max_{a\in\mathcal{A}}\mathrm{Span}\big(r(s,a,\cdot)+\gamma v\big) \\
&\quad\leq \frac{\eta}{2}\big(1+\mathrm{Span}(v)\big).
\end{aligned}$$

This proves the fixed-policy bound with the same joint Wasserstein budget. □

D Empirical DR-DMDP

This appendix develops the bounds required for the empirical DR-DMDP analysis. We first combine the eight perturbation bounds established in Appendix C into a uniform bound for the fixed-policy Bellman operator. We then control $\widehat{v}_\gamma^*-v_\gamma^*$ using the contraction property and relate $\mathrm{Span}(v_\gamma^*)$ to the robust bias span $\mathrm{Span}(u_\delta^*)$.

Lemma 19. *Suppose that $\mathcal{P}$ is a KL-divergence, f_k-divergence, TV, or Wasserstein uncertainty set with either SA-rectangular or S-rectangular structure. If $\eta\leq 1/2$, then, on $\Omega_{n,\eta}$, simultaneously for every $v\in\mathbb{R}^{|\mathcal{S}|}$, $\pi\in\Pi_{\mathrm{SR}}$, and $\gamma\in(0,1)$,*

$$\left\|\widehat{\mathcal{T}}_\gamma^\pi(v)-\mathcal{T}_\gamma^\pi(v)\right\|_\infty \leq 3\eta\big(1+\mathrm{Span}(v)\big).$$

Proof. For the SA-rectangular structure, the KL-, f_k-divergence-, TV-, and Wasserstein-based bounds are given by Corollary 2, Corollary 4, Corollary 6, and Corollary 8, respectively. Their conclusions give, for every $s \in \mathcal{S}$ and $\pi \in \Pi_{\mathrm{SR}}$,

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le \max\left\{3, 2, \frac{1}{2}, \frac{1}{2}\right\} \eta\big(1 + \mathrm{Span}(v)\big).$$

For the S-rectangular structure, the corresponding bounds follow from Corollary 3, Corollary 5, Corollary 7, and Corollary 9, respectively, and yield

$$\left|\widehat{\mathcal{T}}_\gamma^\pi(v)(s) - \mathcal{T}_\gamma^\pi(v)(s)\right| \le \max\left\{3, 2, \frac{1}{2}, \frac{1}{2}\right\} \eta\big(1 + \mathrm{Span}(v)\big).$$

Both displays are bounded by $3\eta\big(1 + \mathrm{Span}(v)\big)$. Taking the maximum over $s \in \mathcal{S}$ proves the claim, simultaneously over v, π, and γ on $\Omega_{n,\eta}$. □

Lemma 20. *Suppose that $\mathcal{P}$ is a KL-divergence, f_k-divergence, TV, or Wasserstein uncertainty set with either SA-rectangular or S-rectangular structure. If $\eta \le 1/2$, then, on $\Omega_{n,\eta}$, for every $\gamma \in (0,1)$,*

$$(1-\gamma)\left\|\widehat{v}_\gamma^* - v_\gamma^*\right\|_\infty \le 3\eta\big(1 + \mathrm{Span}(v_\gamma^*)\big).$$

Proof. Using $\mathcal{T}_\gamma^* = \max_{\pi \in \Pi_{\mathrm{SR}}} \mathcal{T}_\gamma^\pi$ and its empirical counterpart, Lemma 19 gives

$$\left\|\widehat{\mathcal{T}}_\gamma^*(v) - \mathcal{T}_\gamma^*(v)\right\|_\infty \le \sup_{\pi \in \Pi_{\mathrm{SR}}} \left\|\widehat{\mathcal{T}}_\gamma^\pi(v) - \mathcal{T}_\gamma^\pi(v)\right\|_\infty \le 3\eta\big(1 + \mathrm{Span}(v)\big).$$

For the SA-rectangular empirical operator, for $U, v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\begin{aligned}\left|\widehat{\mathcal{T}}_\gamma^*(U)(s) - \widehat{\mathcal{T}}_\gamma^*(v)(s)\right| &\le \gamma \max_{a \in \mathcal{A}} \sup_{q_{s,a} \in \widehat{\mathcal{P}}_{s,a}} |q_{s,a}[U - v]| \\ &\le \gamma \|U - v\|_\infty.\end{aligned}$$

For the S-rectangular empirical operator,

$$\begin{aligned}&\left|\widehat{\mathcal{T}}_\gamma^*(U)(s) - \widehat{\mathcal{T}}_\gamma^*(v)(s)\right| \\ &\le \max_{\phi \in \Delta(\mathcal{A})} \sup_{q_s \in \widehat{\mathcal{P}}_s} \left|\sum_{a \in \mathcal{A}} q_{s,a}[\gamma\phi(a)(U - v)]\right| \\ &\le \gamma \|U - v\|_\infty \max_{\phi \in \Delta(\mathcal{A})} \sup_{q_s \in \widehat{\mathcal{P}}_s} \sum_{a \in \mathcal{A}} \phi(a) q_{s,a}[\mathbf{1}] \\ &= \gamma \|U - v\|_\infty.\end{aligned}$$

Thus, under either structure, $\widehat{\mathcal{T}}_\gamma^*$ is a γ-contraction in $\|\cdot\|_\infty$. Since v_γ^* and $\widehat{v}_\gamma^*$ are the fixed points of $\mathcal{T}_\gamma^*$ in (3.1) and $\widehat{\mathcal{T}}_\gamma^*$, respectively,

$$\begin{aligned}\left\|\widehat{v}_\gamma^* - v_\gamma^*\right\|_\infty &\le \left\|\widehat{\mathcal{T}}_\gamma^*(\widehat{v}_\gamma^*) - \widehat{\mathcal{T}}_\gamma^*(v_\gamma^*)\right\|_\infty \\ &\quad + \left\|\widehat{\mathcal{T}}_\gamma^*(v_\gamma^*) - \mathcal{T}_\gamma^*(v_\gamma^*)\right\|_\infty \\ &\le \gamma \left\|\widehat{v}_\gamma^* - v_\gamma^*\right\|_\infty + 3\eta\big(1 + \mathrm{Span}(v_\gamma^*)\big),\end{aligned}$$

where the last inequality follows from Lemma 19. Rearranging proves the claim. □

Lemma 21. *Suppose that (u_δ^*, g_δ^*) solves the optimal average-reward Bellman equation (3.3) with constant gain under either the SA-rectangular or S-rectangular structure:*

$$\mathcal{T}^*(u_\delta^*) = u_\delta^* + g_\delta^* \mathbf{1}.$$

Then, for every $\gamma \in (0,1)$,

$$\mathrm{Span}(v_\gamma^*) \le 2\,\mathrm{Span}(u_\delta^*).$$

Proof. This extends the comparison argument in the proof of Theorem 4, Appendix B, (2) ⇒ (1), of Wang & Si (2025) from $r(s,a)$ to $r(s,a,s')$, retaining the reward inside the next-state expectation. By additive homogeneity, normalize u_δ^* so that

$$\min_{s\in\mathcal{S}} u_\delta^*(s) = 0, \qquad 0 \le u_\delta^*(s) \le \operatorname{Span}(u_\delta^*).$$

Define

$$\overline{v}_\gamma := \frac{g_\delta^*}{1-\gamma}\mathbf{1} + u_\delta^*, \qquad \underline{v}_\gamma := \overline{v}_\gamma - \operatorname{Span}(u_\delta^*)\mathbf{1}.$$

Since $\gamma u_\delta^* \le u_\delta^*$, the operator definitions (3.1)–(3.2) and the constant-gain equation (3.3) give

$$\begin{aligned}\mathcal{T}_\gamma^*(\overline{v}_\gamma) &= \frac{\gamma g_\delta^*}{1-\gamma}\mathbf{1} + \mathcal{T}_\gamma^*(u_\delta^*)\\ &\le \frac{\gamma g_\delta^*}{1-\gamma}\mathbf{1} + \mathcal{T}^*(u_\delta^*)\\ &= \overline{v}_\gamma.\end{aligned}$$

Also,

$$\gamma u_\delta^* + (1-\gamma)\operatorname{Span}(u_\delta^*)\mathbf{1} \ge u_\delta^*,$$

so

$$\begin{aligned}\mathcal{T}_\gamma^*(\underline{v}_\gamma) &= \left(\frac{\gamma g_\delta^*}{1-\gamma} - \gamma\operatorname{Span}(u_\delta^*)\right)\mathbf{1} + \mathcal{T}_\gamma^*(u_\delta^*)\\ &\ge \left(\frac{\gamma g_\delta^*}{1-\gamma} - \gamma\operatorname{Span}(u_\delta^*)\right)\mathbf{1} + \mathcal{T}^*(u_\delta^*) - (1-\gamma)\operatorname{Span}(u_\delta^*)\mathbf{1}\\ &= \underline{v}_\gamma.\end{aligned}$$

$$\underline{v}_\gamma \le \mathcal{T}_\gamma^*(\underline{v}_\gamma) \le (\mathcal{T}_\gamma^*)^2(\underline{v}_\gamma) \le \cdots \le (\mathcal{T}_\gamma^*)^k(\underline{v}_\gamma),$$
$$\overline{v}_\gamma \ge \mathcal{T}_\gamma^*(\overline{v}_\gamma) \ge (\mathcal{T}_\gamma^*)^2(\overline{v}_\gamma) \ge \cdots \ge (\mathcal{T}_\gamma^*)^k(\overline{v}_\gamma).$$

The γ-contraction property of the discounted operator in (3.1) gives

$$\left\|(\mathcal{T}_\gamma^*)^k(\underline{v}_\gamma) - v_\gamma^*\right\|_\infty \le \gamma^k \left\|\underline{v}_\gamma - v_\gamma^*\right\|_\infty \xrightarrow[k\to\infty]{} 0,$$

$$\left\|(\mathcal{T}_\gamma^*)^k(\overline{v}_\gamma) - v_\gamma^*\right\|_\infty \le \gamma^k \left\|\overline{v}_\gamma - v_\gamma^*\right\|_\infty \xrightarrow[k\to\infty]{} 0.$$

Thus,

$$\begin{aligned}\underline{v}_\gamma \le v_\gamma^* \le \overline{v}_\gamma.&\\ \operatorname{Span}(v_\gamma^*) &\le \max_{s\in\mathcal{S}} \overline{v}_\gamma(s) - \min_{s\in\mathcal{S}} \underline{v}_\gamma(s)\\ &= 2\operatorname{Span}(u_\delta^*).\end{aligned}$$

□

Remark. The constant-gain Bellman solution established in Section 4.1 allows Lemma 21 to bound the discounted value span by twice the robust optimal bias span. Since g_δ^* is the same for every state, the common term $g_\delta^*/(1-\gamma)$ in the upper and lower bounds on v_γ^* cancels in the span estimate, yielding a bound uniform over $\gamma \in (0,1)$. Using this comparison in Lemma 20 gives the dependence on $\operatorname{Span}(u_\delta^*)$ in the estimation bound of Lemma 27, without requiring uniform mixing. The same comparison is used in Lemma 25 to control the discounted-to-average-reward approximation error for Theorem 1 in terms of the same robust optimal bias span.

E Empirical DR-AMDP

This appendix connects the optimal discounted value v_γ^* with the optimal constant gain g_δ^*. The main result is Lemma 25, which bounds $\left\|(1-\gamma)v_\gamma^* - g_\delta^*\mathbf{1}\right\|_\infty$ in terms of $(1-\gamma)\operatorname{Span}(u_\delta^*)$. For completeness, we first establish three intermediate results. Lemma 22 bounds $\mathcal{T}^*(v_\gamma^*) - v_\gamma^*$ using the minimum and maximum of v_γ^*. Lemma 23 shows the property of $\mathcal{T}^*$ in (3.2), and Lemma 24 bounds g_δ^* by the minimum and maximum of $\mathcal{T}^*(v) - v$. Together with Lemma 21, these results yield the bound in Lemma 25.

Lemma 22. *Let*

$$m := \min_{s\in\mathcal{S}} v_\gamma^*(s), \qquad M := \max_{s\in\mathcal{S}} v_\gamma^*(s).$$

Under either the SA-rectangular or S-rectangular structure,

$$v_\gamma^* + (1-\gamma)m\mathbf{1} \le \mathcal{T}^*(v_\gamma^*) \le v_\gamma^* + (1-\gamma)M\mathbf{1}.$$

Proof. For the SA-rectangular structure, for every $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $q_{s,a} \in \mathcal{P}_{s,a}$,

$$m \le q_{s,a}[v_\gamma^*] \le M.$$

Hence

$$\begin{aligned} q_{s,a}[r(s,a,\cdot) + \gamma v_\gamma^*] + (1-\gamma)m &\le q_{s,a}[r(s,a,\cdot) + v_\gamma^*] \\ &\le q_{s,a}[r(s,a,\cdot) + \gamma v_\gamma^*] + (1-\gamma)M. \end{aligned}$$

Taking the infimum over $q_{s,a} \in \mathcal{P}_{s,a}$ and then the maximum over $a \in \mathcal{A}$ yields, by the SA-rectangular reductions in Definition 4,

$$\mathcal{T}_\gamma^*(v_\gamma^*)(s) + (1-\gamma)m \le \mathcal{T}^*(v_\gamma^*)(s) \le \mathcal{T}_\gamma^*(v_\gamma^*)(s) + (1-\gamma)M.$$

For the S-rectangular structure, for every $\phi \in \Delta(\mathcal{A})$ and $q_s \in \mathcal{P}_s$,

$$m \le \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)v_\gamma^*] \le M.$$

Consequently,

$$\begin{aligned} &\sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + \gamma v_\gamma^*)] + (1-\gamma)m \\ &\quad\le \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + v_\gamma^*)] \\ &\quad\le \sum_{a\in\mathcal{A}} q_{s,a}[\phi(a)(r(s,a,\cdot) + \gamma v_\gamma^*)] + (1-\gamma)M. \end{aligned}$$

Taking the infimum over $q_s \in \mathcal{P}_s$ and the maximum over $\phi \in \Delta(\mathcal{A})$ gives the same display using the operator definitions (3.1) and (3.2). Since $v_\gamma^* = \mathcal{T}_\gamma^*(v_\gamma^*)$, the claim follows. □

Lemma 23. *Let $v \in \mathbb{R}^{|\mathcal{S}|}$. Under either the SA-rectangular or S-rectangular structure, if*

$$C_1\mathbf{1} \le \mathcal{T}^*(v) - v \le C_2\mathbf{1},$$

then, for every $n \in \mathbb{N}$,

$$nC_1\mathbf{1} \le \left(\mathcal{T}^*\right)^n(v) - v \le nC_2\mathbf{1}.$$

Proof. Under both structures, the operator definition (3.2) gives

$$U \le W \implies \mathcal{T}^*(U) \le \mathcal{T}^*(W), \qquad \mathcal{T}^*(W + c\mathbf{1}) = \mathcal{T}^*(W) + c\mathbf{1}.$$

The term $r(s,a,\cdot)$ is unchanged in both relations. The assertion holds for $n = 1$. If it holds for n, then

$$\begin{aligned} \left(\mathcal{T}^*\right)^{n+1}(v) &= \mathcal{T}^*\left(\left(\mathcal{T}^*\right)^n(v)\right) \\ &\le \mathcal{T}^*(v + nC_2\mathbf{1}) \\ &= \mathcal{T}^*(v) + nC_2\mathbf{1} \\ &\le v + (n+1)C_2\mathbf{1}, \end{aligned}$$

and

$$\begin{aligned} \left(\mathcal{T}^*\right)^{n+1}(v) &= \mathcal{T}^*\left(\left(\mathcal{T}^*\right)^n(v)\right) \\ &\ge \mathcal{T}^*(v + nC_1\mathbf{1}) \\ &= \mathcal{T}^*(v) + nC_1\mathbf{1} \\ &\ge v + (n+1)C_1\mathbf{1}. \end{aligned}$$

Induction completes the proof. □

Lemma 24. *Suppose that (u_δ^*, g_δ^*) solves the optimal average-reward Bellman equation (3.3) with constant gain under either the SA-rectangular or S-rectangular structure:*

$$\mathcal{T}^*(u_\delta^*) = u_\delta^* + g_\delta^* \mathbf{1}.$$

Then, for every $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\min_{s\in\mathcal{S}} \{\mathcal{T}^*(v)(s) - v(s)\} \le g_\delta^* \le \max_{s\in\mathcal{S}} \{\mathcal{T}^*(v)(s) - v(s)\}.$$

Proof. Let

$$C_1 := \min_{s\in\mathcal{S}} \{\mathcal{T}^*(v)(s) - v(s)\}, \qquad C_2 := \max_{s\in\mathcal{S}} \{\mathcal{T}^*(v)(s) - v(s)\}.$$

By Lemma 23,

$$nC_1\mathbf{1} \le \big(\mathcal{T}^*\big)^n(v) - v \le nC_2\mathbf{1}.$$

Set

$$B_1 := \min_{s\in\mathcal{S}}\big(v(s) - u_\delta^*(s)\big), \qquad B_2 := \max_{s\in\mathcal{S}}\big(v(s) - u_\delta^*(s)\big).$$

Then, using the constant-gain equation (3.3),

$$u_\delta^* + B_1\mathbf{1} \le v \le u_\delta^* + B_2\mathbf{1}, \qquad \big(\mathcal{T}^*\big)^n(u_\delta^*) = u_\delta^* + ng_\delta^*\mathbf{1}.$$

By monotonicity and additive homogeneity,

$$u_\delta^* + (ng_\delta^* + B_1)\mathbf{1} \le \big(\mathcal{T}^*\big)^n(v) \le u_\delta^* + (ng_\delta^* + B_2)\mathbf{1}.$$

Thus,

$$ng_\delta^*\mathbf{1} - (B_2 - B_1)\mathbf{1} \le \big(\mathcal{T}^*\big)^n(v) - v \le ng_\delta^*\mathbf{1} + (B_2 - B_1)\mathbf{1}.$$

Combining the two displays gives

$$C_1 \le g_\delta^* + \frac{B_2 - B_1}{n}, \qquad g_\delta^* - \frac{B_2 - B_1}{n} \le C_2.$$

Letting $n \to \infty$ proves the claim. □

Lemma 25. *Suppose that (u_δ^*, g_δ^*) solves the optimal average-reward Bellman equation (3.3) with constant gain under either the SA-rectangular or S-rectangular structure. Then, for every $\gamma \in (0,1)$,*

$$\big\|(1-\gamma)v_\gamma^* - g_\delta^*\mathbf{1}\big\|_\infty \le 2(1-\gamma)\,\mathrm{Span}(u_\delta^*).$$

Proof. Let

$$m := \min_{s\in\mathcal{S}} v_\gamma^*(s), \qquad M := \max_{s\in\mathcal{S}} v_\gamma^*(s).$$

By Lemma 22 and Lemma 24,

$$(1-\gamma)m \le \min_{s\in\mathcal{S}} \big\{\mathcal{T}^*(v_\gamma^*)(s) - v_\gamma^*(s)\big\} \le g_\delta^* \le \max_{s\in\mathcal{S}} \big\{\mathcal{T}^*(v_\gamma^*)(s) - v_\gamma^*(s)\big\} \le (1-\gamma)M.$$

Thus,

$$m \le \frac{g_\delta^*}{1-\gamma} \le M,$$

and hence

$$\begin{aligned}\big\|(1-\gamma)v_\gamma^* - g_\delta^*\mathbf{1}\big\|_\infty &= (1-\gamma)\max_{s\in\mathcal{S}}\left|v_\gamma^*(s) - \frac{g_\delta^*}{1-\gamma}\right| \\ &\le (1-\gamma)(M-m) \\ &= (1-\gamma)\,\mathrm{Span}(v_\gamma^*).\end{aligned}$$

Finally, Lemma 21 gives $\mathrm{Span}(v_\gamma^*) \le 2\,\mathrm{Span}(u_\delta^*)$; combining this with the preceding display yields the claimed bound. □

The following comparison relates the discounted value to a constant-gain Bellman solution.

Lemma 26. *Suppose that (u_δ^*, g_δ^*) solves the optimal average-reward Bellman equation (3.3) with constant gain under either rectangularity structure. Then, for every $\gamma \in (0,1)$,*

$$\operatorname{Span}(v_\gamma^*) \le 2\operatorname{Span}(u_\delta^*), \qquad \left\|(1-\gamma)v_\gamma^* - g_\delta^*\mathbf{1}\right\|_\infty \le 2(1-\gamma)\operatorname{Span}(u_\delta^*).$$

These bounds follow from the comparison argument in the proof of Theorem 4, Appendix B, (2) $\Rightarrow$ (1), of Wang & Si (2025). We extend that argument to bounded transition-dependent rewards $r(s,a,s')$ by retaining the reward inside the next-state expectation.

Proof. The first inequality is the discounted-value span bound proved in Lemma 21; the second is Lemma 25 above. Both hold for every $\gamma \in (0,1)$ under either rectangularity structure. □

F Main Results

We first establish an auxiliary discounted-value bound and use it to prove Theorem 1 from Section 4. We then prove the explicit bias-span bounds in Theorem 2.

Lemma 27. *Suppose either Assumption 1 or Assumption 2 holds, and let (u_δ^*, g_δ^*) be any solution of the corresponding optimal average-reward Bellman equation (3.3) with constant gain and $\operatorname{Span}(u_\delta^*) \ge 1$. Fix $\beta \in (0,1)$ and $\gamma \in (0,1)$. If*

$$n \ge \frac{16}{p_\wedge}\log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right),$$

then the empirical optimal discounted value $\widehat{v}_\gamma^*$ satisfies, with probability at least $1-\beta$,

$$\left\|\widehat{v}_\gamma^* - v_\gamma^*\right\|_\infty \le \frac{14}{1-\gamma}\sqrt{\frac{1}{np_\wedge}\log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)}\operatorname{Span}(u_\delta^*).$$

Proof of Lemma 27. Let η be chosen as in Proposition 3. The sample-size condition implies

$$\sqrt{\frac{1}{np_\wedge}\log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)} \le \frac{1}{4},$$

and hence

$$\eta \le \left(\frac{1}{12}+\sqrt{2}\right)\sqrt{\frac{1}{np_\wedge}\log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)} < \frac{1}{2}.$$

On $\Omega_{n,\eta}$, Lemma 20 and Lemma 21 give

$$\begin{aligned}(1-\gamma)\left\|\widehat{v}_\gamma^* - v_\gamma^*\right\|_\infty &\le 3\eta\big(1+\operatorname{Span}(v_\gamma^*)\big)\\ &\le 3\eta\big(1+2\operatorname{Span}(u_\delta^*)\big)\\ &\le 9\eta\operatorname{Span}(u_\delta^*)\\ &\le 14\sqrt{\frac{1}{np_\wedge}\log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)}\operatorname{Span}(u_\delta^*).\end{aligned}$$

The penultimate inequality uses $\operatorname{Span}(u_\delta^*) \ge 1$, and the last uses $9(1/12+\sqrt{2}) < 14$. Since $\mathbb{P}(\Omega_{n,\eta}) \ge 1-\beta$ by Proposition 3, dividing by $1-\gamma$ proves the claim. □

Remark. Lemma 27 implies that the empirical optimal discounted value $\widehat{v}_\gamma^*$ is a uniform ϵ-error estimate of v_γ^* with high probability using $\widetilde{\mathcal{O}}\big(|\mathcal{S}||\mathcal{A}|p_\wedge^{-1}(1-\gamma)^{-2}\operatorname{Span}^2(u_\delta^*)\epsilon^{-2}\big)$ samples. The proof separates the optimal-value error into a contraction term and an operator-perturbation term. The Bellman fixed-point identities and the triangle inequality give $\|\widehat{v}_\gamma^* - v_\gamma^*\|_\infty \le \|\widehat{\mathcal{T}}_\gamma^*(\widehat{v}_\gamma^*) - \widehat{\mathcal{T}}_\gamma^*(v_\gamma^*)\|_\infty + \|\widehat{\mathcal{T}}_\gamma^*(v_\gamma^*) - \mathcal{T}_\gamma^*(v_\gamma^*)\|_\infty$. The first term is at most $\gamma\|\widehat{v}_\gamma^* - v_\gamma^*\|_\infty$ by contraction; the second measures the operator perturbation at the population value and is at most $\sup_{\pi\in\Pi_{\mathrm{SR}}}\|\widehat{\mathcal{T}}_\gamma^\pi(v_\gamma^*) - \mathcal{T}_\gamma^\pi(v_\gamma^*)\|_\infty$ by $\mathcal{T}_\gamma^* = \max_{\pi\in\Pi_{\mathrm{SR}}}\mathcal{T}_\gamma^\pi$ and its empirical counterpart. This is controlled by Corollaries 2–9. Building on the constant-gain Bellman framework of Wang & Si (2025), we use

$\mathrm{Span}(v_\gamma^*) \le 2\,\mathrm{Span}(u_\delta^*)$ to express the bound in terms of the robust optimal bias span. The required Bellman structure follows from nominal weak communication and the stated radius restrictions for divergence-based sets, or from a positive radius for metric-based sets. This gives the $(1-\gamma)^{-2}$ sample-complexity dependence and the $n^{-1/2}$ robust gain-estimation rate in Theorem 1.

Proof of Theorem 1. Let η be chosen as in Proposition 3. The sample-size condition gives $\eta < 1/2$, as shown in the proof of Lemma 27. Since $1-\gamma_n = 1/\sqrt{n}$,

$$\left\| \frac{1}{\sqrt{n}} \widehat{v}^*_{\gamma_n} - g^*_\delta \mathbf{1} \right\|_\infty \le (1-\gamma_n) \left\| \widehat{v}^*_{\gamma_n} - v^*_{\gamma_n} \right\|_\infty + \left\| (1-\gamma_n) v^*_{\gamma_n} - g^*_\delta \mathbf{1} \right\|_\infty .$$

By Lemma 27 and Lemma 25, on $\Omega_{n,\eta}$,

$$\left\| \frac{1}{\sqrt{n}} \widehat{v}^*_{\gamma_n} - g^*_\delta \mathbf{1} \right\|_\infty \le 14 \sqrt{\frac{1}{np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta} \right)} \mathrm{Span}(u^*_\delta) + \frac{2}{\sqrt{n}} \mathrm{Span}(u^*_\delta).$$

The assumption $\mathrm{Span}(u^*_\delta) \ge 1$ implies $|\mathcal{S}| \ge 2$, so

$$\frac{1}{\sqrt{n}} \le \sqrt{\frac{1}{np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta} \right)}.$$

Therefore,

$$\frac{2}{\sqrt{n}} \mathrm{Span}(u^*_\delta) \le 2 \sqrt{\frac{1}{np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta} \right)} \mathrm{Span}(u^*_\delta).$$

Thus,

$$\left\| \frac{1}{\sqrt{n}} \widehat{v}^*_{\gamma_n} - g^*_\delta \mathbf{1} \right\|_\infty \le 16 \sqrt{\frac{1}{np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta} \right)} \mathrm{Span}(u^*_\delta).$$

We now bound the robust average reward of the returned policy on the same event $\Omega_{n,\eta}$. By construction in Algorithm 1, $\widehat{\pi}^*$ attains the maximization in $\widehat{\mathcal{T}}^*_{\gamma_n}$ at $\widehat{v}^*_{\gamma_n}$, so $\widehat{\mathcal{T}}^{\widehat{\pi}^*}_{\gamma_n}(\widehat{v}^*_{\gamma_n}) = \widehat{\mathcal{T}}^*_{\gamma_n}(\widehat{v}^*_{\gamma_n}) = \widehat{v}^*_{\gamma_n}$. Since $\widehat{\mathcal{T}}^{\widehat{\pi}^*}_{\gamma_n}$ is a γ_n-contraction, its unique fixed point is the robust discounted value of $\widehat{\pi}^*$; hence $\widehat{\pi}^*$ is robust discount-optimal in the empirical MDP. This is the standard optimality characterization under SA- and S-rectangularity; see Iyengar (2005) and Wiesemann et al. (2013), respectively. Lemma 19 therefore implies, componentwise,

$$\begin{aligned} \mathcal{T}^{\widehat{\pi}^*}_{\gamma_n}(\widehat{v}^*_{\gamma_n}) &\ge \widehat{\mathcal{T}}^{\widehat{\pi}^*}_{\gamma_n}(\widehat{v}^*_{\gamma_n}) - 3\eta\big(1 + \mathrm{Span}(\widehat{v}^*_{\gamma_n})\big)\mathbf{1} \\ &= \widehat{v}^*_{\gamma_n} - 3\eta\big(1 + \mathrm{Span}(\widehat{v}^*_{\gamma_n})\big)\mathbf{1}. \end{aligned}$$

The bound is simultaneous over policies and value vectors, so it applies to the sample-dependent $\widehat{\pi}^*$ and $\widehat{v}^*_{\gamma_n}$ without an additional probability event.

Fix the samples in $\Omega_{n,\eta}$, any $\mathbf{Q} \in (\mathcal{P})^{\mathbb{N}}$, and an initial state $S_0 = s$. Since $Q_t \in \mathcal{P}$ at each time t, the definition of the fixed-policy operator gives

$$\begin{aligned} &\mathrm{E}^{\widehat{\pi}^*}_{\mathbf{Q}} \left[r(S_t, A_t, S_{t+1}) + \gamma_n \widehat{v}^*_{\gamma_n}(S_{t+1}) \mid S_t \right] \\ &\quad \ge \mathcal{T}^{\widehat{\pi}^*}_{\gamma_n}(\widehat{v}^*_{\gamma_n})(S_t) \\ &\quad \ge \widehat{v}^*_{\gamma_n}(S_t) - 3\eta\big(1 + \mathrm{Span}(\widehat{v}^*_{\gamma_n})\big). \end{aligned}$$

Taking expectations and summing over $t = 0, \ldots, T-1$, we use the identity

$$\begin{aligned} &\sum_{t=0}^{T-1} \left[\widehat{v}^*_{\gamma_n}(S_t) - \gamma_n \widehat{v}^*_{\gamma_n}(S_{t+1}) \right] \\ &\quad = \frac{1}{\sqrt{n}} \sum_{t=0}^{T-1} \widehat{v}^*_{\gamma_n}(S_t) + \gamma_n \left[\widehat{v}^*_{\gamma_n}(S_0) - \widehat{v}^*_{\gamma_n}(S_T) \right]. \end{aligned}$$

Dividing by T, bounding each value by its minimum, and bounding the terminal difference by minus the span yields

$$\frac{1}{T}\mathrm{E}_{\mathbf{Q}}^{\widehat{\pi}^*}\left[\sum_{t=0}^{T-1} r(S_t, A_t, S_{t+1}) \,\middle|\, S_0 = s\right] \geq \frac{1}{\sqrt{n}} \min_{x\in\mathcal{S}} \widehat{v}^*_{\gamma_n}(x) - 3\eta\big(1 + \mathrm{Span}(\widehat{v}^*_{\gamma_n})\big) - \frac{\gamma_n}{T}\,\mathrm{Span}(\widehat{v}^*_{\gamma_n}).$$

For fixed samples, $\widehat{v}^*_{\gamma_n}$ is finite. Taking the $\limsup$ as $T \to \infty$ for each $\mathbf{Q}$, and then taking the infimum over $\mathbf{Q} \in (\mathcal{P})^{\mathbb{N}}$, gives

$$g_\delta^{\widehat{\pi}^*}(s) \geq \frac{1}{\sqrt{n}} \min_{x\in\mathcal{S}} \widehat{v}^*_{\gamma_n}(x) - 3\eta\big(1 + \mathrm{Span}(\widehat{v}^*_{\gamma_n})\big).$$

This uses the original long-run criterion without interchanging the infimum and the limit or requiring a constant gain for $\widehat{\pi}^*$. By optimality of g^*_δ, adding and subtracting the same estimate gives

$$\begin{aligned} 0 \leq g^*_\delta - g_\delta^{\widehat{\pi}^*}(s) &= g^*_\delta - \frac{1}{\sqrt{n}} \min_{x\in\mathcal{S}} \widehat{v}^*_{\gamma_n}(x) \\ &\quad + \frac{1}{\sqrt{n}} \min_{x\in\mathcal{S}} \widehat{v}^*_{\gamma_n}(x) - g_\delta^{\widehat{\pi}^*}(s) \\ &\leq \left\| g^*_\delta \mathbf{1} - \frac{\widehat{v}^*_{\gamma_n}}{\sqrt{n}} \right\|_\infty + 3\eta\big(1 + \mathrm{Span}(\widehat{v}^*_{\gamma_n})\big). \end{aligned}$$

By Lemma 20, Lemma 21, and $\mathrm{Span}(u^*_\delta) \geq 1$,

$$\begin{aligned} \frac{1}{\sqrt{n}} \|\widehat{v}^*_{\gamma_n} - v^*_{\gamma_n}\|_\infty &\leq 3\eta\big(1 + \mathrm{Span}(v^*_{\gamma_n})\big) \\ &\leq 3\eta\big(1 + 2\,\mathrm{Span}(u^*_\delta)\big) \\ &\leq 9\eta\,\mathrm{Span}(u^*_\delta). \end{aligned}$$

Consequently,

$$\begin{aligned} \mathrm{Span}(\widehat{v}^*_{\gamma_n}) &\leq \mathrm{Span}(v^*_{\gamma_n}) + 2\|\widehat{v}^*_{\gamma_n} - v^*_{\gamma_n}\|_\infty \\ &\leq \big(2 + 18\sqrt{n}\,\eta\big)\,\mathrm{Span}(u^*_\delta). \end{aligned}$$

The triangle inequality and Lemma 25 also give

$$\left\| g^*_\delta \mathbf{1} - \frac{\widehat{v}^*_{\gamma_n}}{\sqrt{n}} \right\|_\infty \leq \left(9\eta + \frac{2}{\sqrt{n}}\right) \mathrm{Span}(u^*_\delta).$$

Combining these bounds and using $\mathrm{Span}(u^*_\delta) \geq 1$, we obtain

$$\begin{aligned} \|g^*_\delta \mathbf{1} - g_\delta^{\widehat{\pi}^*}\|_\infty &\leq \left(9\eta + \frac{2}{\sqrt{n}}\right) \mathrm{Span}(u^*_\delta) \\ &\quad + 3\eta\left[1 + \big(2 + 18\sqrt{n}\,\eta\big)\,\mathrm{Span}(u^*_\delta)\right] \\ &\leq \left(\frac{2}{\sqrt{n}} + 18\eta + 54\sqrt{n}\,\eta^2\right) \mathrm{Span}(u^*_\delta). \end{aligned}$$

For the choice in Proposition 3, the sample-size condition gives

$$\begin{aligned} \eta &= \frac{1}{3np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right) + \sqrt{\frac{2}{np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)} \\ &\leq \left(\frac{1}{12} + \sqrt{2}\right) \sqrt{\frac{1}{np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)} \\ &\leq \frac{3}{2} \sqrt{\frac{1}{np_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)} \leq \frac{3}{8}. \end{aligned}$$

Substituting this estimate yields

$$\|g_\delta^* \mathbf{1} - g_\delta^{\widehat{\pi}^*}\|_\infty \le \frac{\mathrm{Span}(u_\delta^*)}{\sqrt{n}} \left[2 + 27\sqrt{\frac{1}{p_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right)} + \frac{243}{2p_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right) \right].$$

Since $\mathrm{Span}(u_\delta^*) \ge 1$ implies $|\mathcal{S}| \ge 2$, and $p_\wedge \le 1$ and $\beta \in (0,1)$,

$$\frac{1}{p_\wedge} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right) \ge 1.$$

Thus both the constant and square-root terms are bounded by the same logarithmic term, and $2 + 27 + 243/2 = 301/2 \le 151$ gives

$$\|g_\delta^* \mathbf{1} - g_\delta^{\widehat{\pi}^*}\|_\infty \le \frac{151\ \mathrm{Span}(u_\delta^*)}{p_\wedge \sqrt{n}} \log\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|}{\beta}\right).$$

Both inequalities hold on the same event $\Omega_{n,\eta}$, whose probability is at least $1 - \beta$ by Proposition 3. This proves the claim. □

Metric-based and divergence-based uncertainty sets require separate analyses because they have different support properties. A positive-radius TV or Wasserstein uncertainty set can move probability mass outside the nominal support. We first prove the metric bounds.

Proof of Theorem 2. Fix $\gamma \in (0,1)$, write $v = v_\gamma^*$, and let

$$s_+ \in \arg\max_{s \in \mathcal{S}} v(s), \qquad s_- \in \arg\min_{s \in \mathcal{S}} v(s).$$

Since $r(s,a,s') \in [0,1]$, the SA-rectangular and S-rectangular operators in (3.1) both give

$$v(s_-) = \mathcal{T}_\gamma^*(v)(s_-) \ge \gamma v(s_-), \qquad (1-\gamma) v(s_-) \ge 0.$$

Case 1: TV uncertainty sets. The uncertainty sets are compact and the Bellman objectives in Definition 4 are continuous; hence the outer maximum and inner infimum are attained. Choose $a_+ \in \mathcal{A}$ and $\phi_+ \in \Delta(\mathcal{A})$ attaining the respective SA-rectangular and S-rectangular maxima at s_+. Set

$$q_{s_+,a} = (1-\delta) p_{s_+,a} + \delta \mathbf{1}\{\cdot = s_-\}.$$

Then, for every $a \in \mathcal{A}$,

$$\begin{aligned} d_{\mathrm{TV}}(q_{s_+,a}, p_{s_+,a}) &= \frac{\delta}{2} \left\| \mathbf{1}\{\cdot = s_-\} - p_{s_+,a} \right\|_1 \\ &= \delta\big(1 - p_{s_+,a}(s_-)\big) \le \delta, \\ q_{s_+,a}[v] &= (1-\delta) p_{s_+,a}[v] + \delta v(s_-) \\ &\le (1-\delta) v(s_+) + \delta v(s_-). \end{aligned}$$

Since $q_{s_+,a_+} \in \mathcal{P}_{s_+,a_+}$, the SA-rectangular reduction in Definition 4 gives

$$\begin{aligned} v(s_+) &= \inf_{q_{s_+,a_+} \in \mathcal{P}_{s_+,a_+}} q_{s_+,a_+}[r(s_+,a_+,\cdot) + \gamma v] \\ &\le q_{s_+,a_+}[r(s_+,a_+,\cdot) + \gamma v] \\ &\le 1 + \gamma\big((1-\delta) v(s_+) + \delta v(s_-)\big). \end{aligned}$$

For the S-rectangular structure,

$$\sum_{a \in \mathcal{A}} d_{\mathrm{TV}}(q_{s_+,a}, p_{s_+,a}) \le |\mathcal{A}| \delta,$$

so $(q_{s_+,a})_{a\in\mathcal{A}} \in \mathcal{P}_{s_+}$. Therefore, the operator definition (3.1) gives

$$\begin{aligned} v(s_+) &= \inf_{q_s\in\mathcal{P}_{s_+}} \sum_{a\in\mathcal{A}} q_{s,a}\left[\phi_+(a)(r(s_+,a,\cdot)+\gamma v)\right] \\ &\le \sum_{a\in\mathcal{A}} q_{s_+,a}\left[\phi_+(a)(r(s_+,a,\cdot)+\gamma v)\right] \\ &\le 1+\gamma\sum_{a\in\mathcal{A}}\phi_+(a)q_{s_+,a}[v] \\ &\le 1+\gamma\big((1-\delta)v(s_+)+\delta v(s_-)\big). \end{aligned}$$

Thus, under either structure,

$$\begin{aligned} \operatorname{Span}(v^*_\gamma) &= v(s_+)-v(s_-) \\ &\le 1+\gamma(1-\delta)v(s_+)+(\gamma\delta-1)v(s_-) \\ &= 1+\gamma(1-\delta)\operatorname{Span}(v^*_\gamma)-(1-\gamma)v(s_-) \\ &\le 1+\gamma(1-\delta)\operatorname{Span}(v^*_\gamma), \end{aligned}$$

and hence

$$\operatorname{Span}(v^*_\gamma) \le \frac{1}{1-\gamma(1-\delta)} \le \frac{1}{\delta}.$$

Case 2: Wasserstein uncertainty sets. Again, the uncertainty sets are compact and the Bellman objectives in Definition 4 are continuous, so the outer maximum and inner infimum are attained. Choose a_+ and ϕ_+ attaining the respective maxima at s_+. Let

$$\alpha = \left(\frac{c}{\delta}\right)^{-\ell}, \qquad q_{s_+,a} = (1-\alpha)p_{s_+,a} + \alpha\mathbf{1}\{\cdot = s_-\}.$$

Coupling the mass $(1-\alpha)p_{s_+,a}$ with itself and transporting the remaining mass to s_- gives

$$W_\ell(q_{s_+,a}, p_{s_+,a})^\ell \le \alpha c^\ell = \delta^\ell, \qquad q_{s_+,a}[v] \le (1-\alpha)v(s_+)+\alpha v(s_-).$$

Thus $q_{s_+,a_+} \in \mathcal{P}_{s_+,a_+}$ in the SA-rectangular case and $(q_{s_+,a})_{a\in\mathcal{A}} \in \mathcal{P}_{s_+}$ in the S-rectangular case, since

$$\sum_{a\in\mathcal{A}} W_\ell(q_{s_+,a}, p_{s_+,a})^\ell \le |\mathcal{A}|\delta^\ell.$$

The same SA- and S-rectangular Bellman arguments as in Case 1, with α in place of δ, give

$$v(s_+) \le 1+\gamma\big((1-\alpha)v(s_+)+\alpha v(s_-)\big).$$

Consequently,

$$\operatorname{Span}(v^*_\gamma) \le 1+\gamma(1-\alpha)\operatorname{Span}(v^*_\gamma).$$

Hence

$$\operatorname{Span}(v^*_\gamma) \le \frac{1}{1-\gamma(1-\alpha)} \le \left(\frac{c}{\delta}\right)^\ell.$$

Applying the normalization and convergent-subsequence argument from the proof of Proposition 2 to these uniform bounds, the bias can be chosen so that

$$\operatorname{Span}(u^*_\delta) \le \frac{1}{\delta} \quad \text{for TV}, \qquad \operatorname{Span}(u^*_\delta) \le \left(\frac{c}{\delta}\right)^\ell \quad \text{for Wasserstein}.$$

Case 3: KL and f_k uncertainty sets. For SA-rectangular uncertainty, the coordinatewise arguments of Chen et al. (2025, Proposition B.1) for KL divergence and Chen et al. (2025, Lemma E.1) for f_k divergence show that the stated radius conditions imply

$$p_{s,a}(s') > 0 \quad\Longrightarrow\quad q_{s,a}(s') \ge \frac{1}{2}p_{s,a}(s') \ge \frac{1}{2}p_\wedge.$$

For S-rectangular uncertainty, each nonnegative action-wise divergence is at most $|\mathcal{A}|\delta$, so the same argument applies under the corresponding stated radius. Since $q_{s,a} \ll p_{s,a}$, every feasible row has the same support as its nominal row. Thus every nominal communication path is retained.

Fix an arbitrary solution (u_δ^*, g_δ^*) of (3.3). For SA-rectangular uncertainty, choose a minimizing row $q_{s,a}$ in the SA-rectangular reduction in Definition 4 for every (s,a). For S-rectangular uncertainty, compactness and Lemma 2 give, for every $s \in \mathcal{S}$,

$$q_s \in \arg\min_{q_s \in \mathcal{P}_s} \max_{\phi \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} q_{s,a}\big[\phi(a)\big(r(s,a,\cdot) + u_\delta^*\big)\big]\,.$$

Since the objective is linear in ϕ,

$$\max_{\phi \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} q_{s,a}\big[\phi(a)\big(r(s,a,\cdot) + u_\delta^*\big)\big] = \max_{a \in \mathcal{A}} q_{s,a}[r(s,a,\cdot) + u_\delta^*]\,.$$

Thus, by the average-reward operator definition (3.2), under either structure, the worst-case transition kernel $Q \in \mathcal{P}$ satisfies

$$u_\delta^*(s) + g_\delta^* = \max_{a \in \mathcal{A}} q_{s,a}[r(s,a,\cdot) + u_\delta^*]\,, \qquad s \in \mathcal{S}.$$

By Lemma 4, Q is weakly communicating; let C denote its communicating set. Every positive coordinate of Q is at least $p_\wedge/2$.

Let

$$s^+ \in \arg\max_{s \in \mathcal{S}} u_\delta^*(s), \qquad s^- \in \arg\min_{s \in \mathcal{S}} u_\delta^*(s).$$

The fixed-kernel Bellman equation gives

$$u_\delta^*(s^-) + g_\delta^* \geq u_\delta^*(s^-), \qquad u_\delta^*(s^+) + g_\delta^* \leq u_\delta^*(s^+) + 1,$$

so $0 \leq g_\delta^* \leq 1$.

Let $y \in C$. Weak communication allows us to choose $\pi \in \Pi_{\mathrm{SD}}$ that puts probability one on one action at each state and reaches y almost surely from every state: choosing actions along shortest positive-probability paths to y leaves no closed class outside y. Define

$$\tau_y := \inf\{t \geq 0 : S_t = y\}, \qquad M := \max_{s \in \mathcal{S}} \mathbb{E}_s^{\pi,Q}[\tau_y].$$

Choose s_0 attaining M and a simple positive-probability path under π,

$$s_0, s_1, \ldots, s_m = y, \qquad m \leq |\mathcal{S}| - 1.$$

For $0 \leq i < m$, take $a \in \mathcal{A}$ with $\pi(a \mid s_i) = 1$. Then

$$\begin{aligned}\mathbb{E}_{s_i}^{\pi,Q}[\tau_y] &\leq 1 + q_{s_i,a}(s_{i+1})\mathbb{E}_{s_{i+1}}^{\pi,Q}[\tau_y] \\ &\quad + \big(1 - q_{s_i,a}(s_{i+1})\big)M,\end{aligned}$$

where $q_{s_i,a}(s_{i+1}) \geq p_\wedge/2$. Rearranging the case $i = 0$ and applying induction along the path gives

$$M \leq \sum_{j=1}^{i} \left(\frac{2}{p_\wedge}\right)^j + \mathbb{E}_{s_i}^{\pi,Q}[\tau_y], \qquad 1 \leq i \leq m.$$

Setting $i = m$ yields

$$\begin{aligned}\max_{s \in \mathcal{S}} \mathbb{E}_s^{\pi,Q}[\tau_y] &\leq \sum_{j=1}^{|\mathcal{S}|-1} \left(\frac{2}{p_\wedge}\right)^j \\ &= \frac{2}{2 - p_\wedge}\left[\left(\frac{2}{p_\wedge}\right)^{|\mathcal{S}|-1} - 1\right].\end{aligned}$$

The same argument bounds the first entrance time of C under any stationary deterministic policy, since states outside C are transient under every such policy; replace the terminal state by the target set C.

Let

$$y^+ \in \arg\max_{y \in C} u_\delta^*(y), \qquad \tau_C := \inf\{t \geq 0 : S_t \in C\}, \qquad \tau_{y^+} := \inf\{t \geq 0 : S_t = y^+\}.$$

Let $\pi^* \in \Pi_{\mathrm{SD}}$ put probability one on a maximizing action in the fixed-kernel Bellman equation at each state, and choose $\pi \in \Pi_{\mathrm{SD}}$ as above with target y^+. Applying the Bellman equality under π^* up to τ_C and the Bellman inequality under π up to τ_{y^+} gives

$$\begin{aligned} u_\delta^*(s^+) &= \mathbb{E}_{s^+}^{\pi^*,Q}\left[\sum_{t=0}^{\tau_C-1}\big(r(S_t,A_t,S_{t+1})-g_\delta^*\big)+u_\delta^*(S_{\tau_C})\right] \\ &\le u_\delta^*(y^+)+(1-g_\delta^*)\mathbb{E}_{s^+}^{\pi^*,Q}[\tau_C], \\ u_\delta^*(s^-) &\ge \mathbb{E}_{s^-}^{\pi,Q}\left[\sum_{t=0}^{\tau_{y^+}-1}\big(r(S_t,A_t,S_{t+1})-g_\delta^*\big)+u_\delta^*(y^+)\right] \\ &\ge u_\delta^*(y^+)-g_\delta^*\mathbb{E}_{s^-}^{\pi,Q}[\tau_{y^+}]. \end{aligned}$$

The stopped relations follow by truncation, since both hitting times have finite expectation. Therefore,

$$\begin{aligned} \mathrm{Span}(u_\delta^*) &\le (1-g_\delta^*)\mathbb{E}_{s^+}^{\pi^*,Q}[\tau_C]+g_\delta^*\mathbb{E}_{s^-}^{\pi,Q}[\tau_{y^+}] \\ &\le \frac{2}{2-p_\wedge}\left[\left(\frac{2}{p_\wedge}\right)^{|\mathcal{S}|-1}-1\right]. \end{aligned}$$

□

Remark. The TV and Wasserstein span bounds are attained. Let $\mathcal{S}=\{s_+,s_-\}$, $\mathcal{A}=\{a\}$, and $r(s_+,a,\cdot)=1$, $r(s_-,a,\cdot)=0$. For TV uncertainty, take

$$p_{s_+,a}=\mathbf{1}\{\cdot=s_+\},\qquad p_{s_-,a}=\delta\mathbf{1}\{\cdot=s_+\}+(1-\delta)\mathbf{1}\{\cdot=s_-\}.$$

At radius δ, the minimizing kernels are

$$q_{s_+,a}=(1-\delta)\mathbf{1}\{\cdot=s_+\}+\delta\mathbf{1}\{\cdot=s_-\},\qquad q_{s_-,a}=\mathbf{1}\{\cdot=s_-\},$$

which yield $g_\delta^*=0$, $u_\delta^*(s_-)=0$, and $u_\delta^*(s_+)=1/\delta$.

For Wasserstein uncertainty, let $\rho(s_+,s_-)=c$, $\alpha=(\delta/c)^\ell$, and take

$$p_{s_+,a}=\mathbf{1}\{\cdot=s_+\},\qquad p_{s_-,a}=\alpha\mathbf{1}\{\cdot=s_+\}+(1-\alpha)\mathbf{1}\{\cdot=s_-\}.$$

The minimizing kernels are

$$q_{s_+,a}=(1-\alpha)\mathbf{1}\{\cdot=s_+\}+\alpha\mathbf{1}\{\cdot=s_-\},\qquad q_{s_-,a}=\mathbf{1}\{\cdot=s_-\},$$

and satisfy $W_\ell(q_{s,a},p_{s,a})^\ell=\alpha c^\ell=\delta^\ell$. Thus $g_\delta^*=0$, $u_\delta^*(s_-)=0$, and

$$u_\delta^*(s_+)=\frac{1}{\alpha}=\left(\frac{c}{\delta}\right)^\ell.$$

Proof of Theorem 3. Fix $0<p_\wedge\le\frac{1}{2}$ and let

$$\mathcal{S}=\{S_1,S_2,\dots,S_{|\mathcal{S}|}\},\qquad \mathcal{A}=\{a\},$$

and define

$$p_{S_1,a}=\mathbf{1}\{\cdot=S_1\},\qquad p_{S_i,a}=p_\wedge\mathbf{1}\{\cdot=S_{i-1}\}+(1-p_\wedge)\mathbf{1}\{\cdot=S_{|\mathcal{S}|}\},\quad 2\le i\le|\mathcal{S}|,$$

with

$$r(S_1,a,\cdot)=1,\qquad r(S_i,a,\cdot)=0,\quad 2\le i\le|\mathcal{S}|.$$

The unique communicating class is $\{S_1\}$, while $S_2,\dots,S_{|\mathcal{S}|}$ are transient under the unique policy. With one action, the SA- and S-rectangular uncertainty sets coincide.

For every sufficiently small positive δ, choose each $q_{S_i,a}$ to minimize the feasible probability assigned to S_{i-1}, for $2\le i\le|\mathcal{S}|$. The nominal rows have the same two probabilities, and KL and f_k divergences are invariant under relabeling, so their minimum feasible predecessor probabilities agree. The support bound from Case 3 and feasibility of the nominal rows give

$$\frac{1}{2}p_\wedge\le q_{S_2,a}(S_1)\le p_\wedge,\qquad q_{S_i,a}(S_{i-1})=q_{S_2,a}(S_1),\quad 2\le i\le|\mathcal{S}|.$$

Normalize $u_\delta^*(S_{|\mathcal{S}|}) = 0$. The constant-gain equation (3.3) at S_1 gives $g_\delta^* = 1$. Starting from $i = |\mathcal{S}|$ and proceeding backward, $u_\delta^*(S_i) \geq 0$ implies

$$1 + u_\delta^*(S_i) = \inf_{q_{S_i,a} \in \mathcal{P}_{S_i,a}} q_{S_i,a}(S_{i-1}) u_\delta^*(S_{i-1}) > 0.$$

Thus $u_\delta^*(S_{i-1}) > 0$, and the infimum is attained by the chosen $q_{S_i,a}$. Hence

$$u_\delta^*(S_{i-1}) = \frac{1 + u_\delta^*(S_i)}{q_{S_i,a}(S_{i-1})}, \qquad 2 \leq i \leq |\mathcal{S}|.$$

Consequently,

$$u_\delta^*(S_i) = \sum_{j=1}^{|\mathcal{S}|-i} \frac{1}{\big(q_{S_2,a}(S_1)\big)^j}, \qquad 1 \leq i \leq |\mathcal{S}| - 1.$$

The constant-gain equation (3.3) at S_1 uniquely determines g_δ^*, and the displayed backward recursion uniquely determines the normalized bias. Thus all solutions differ only by an additive constant and have the same span. Therefore,

$$\begin{aligned} \mathrm{Span}(u_\delta^*) &= \sum_{j=1}^{|\mathcal{S}|-1} \frac{1}{\big(q_{S_2,a}(S_1)\big)^j} \\ &\geq \sum_{j=1}^{|\mathcal{S}|-1} \left(\frac{1}{p_\wedge}\right)^j \\ &= \frac{p_\wedge^{-(|\mathcal{S}|-1)} - 1}{1 - p_\wedge}. \end{aligned}$$

□

G Additional Experiments

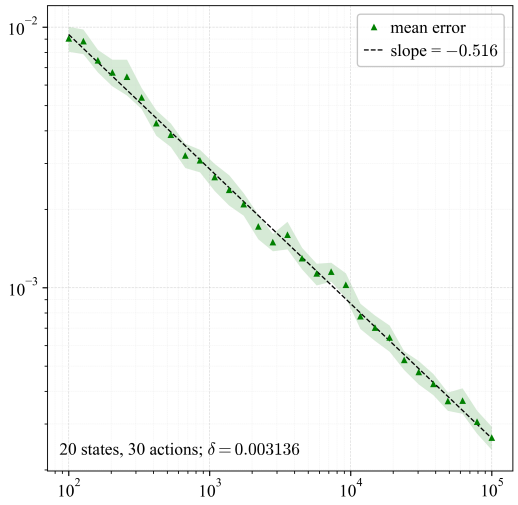
This appendix presents additional large-scale experiments on a fixed nominal MDP with 20 states and 30 actions. We evaluate all eight combinations of SA- and S-rectangularity with KL-divergence, χ^2-divergence, TV, and Wasserstein uncertainty sets.

We construct a random nominal kernel P by partitioning the states into $T = \{0, 1, 2, 3\}$ and $C = C_1 \cup C_2$, where $C_1 = \{4, \ldots, 11\}$ and $C_2 = \{12, \ldots, 19\}$. At every state in T, each action assigns total probability $1/2$ to T and $1/2$ to C. At a state in C_i, actions $a_1, \ldots, a_{15}$ remain in C_i, while actions $a_{16}, \ldots, a_{30}$ can reach every state in C. Each conditional distribution on an allowed set of m states is drawn independently as $0.8z + 0.2\mathbf{1}/m$, with $z \sim \mathrm{Dirichlet}(1, \ldots, 1)$. All other transition probabilities are zero. We generate rewards as $\sigma_{s,a} Z_{s,a,s'}$, where $\sigma_{s,a} \sim \mathrm{Uniform}(0, 1)$ and $Z_{s,a,s'} \sim \mathcal{N}(0, 1)$, and rescale the full reward array to $[0, 1]$. The nominal kernel and rewards remain fixed across all trials.

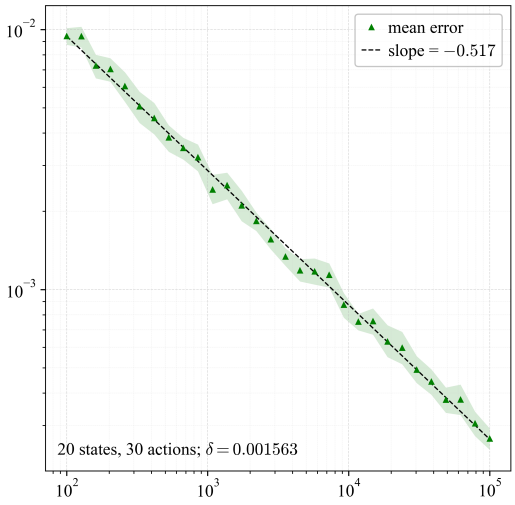
Starting in T, the probability of remaining there after t transitions is 2^{-t} under every policy, and no action returns from C to T. Thus every state in T is transient under every policy. The stationary policy that always chooses a_{16} makes all states in C communicate, so the nominal MDP is weakly communicating in the sense of Definition 1. The stationary policy that always chooses a_1 has two disjoint recurrent classes, C_1 and C_2, so the nominal MDP is not unichain. Its induced Markov chain cannot converge to a single stationary distribution independent of the initial state. Since $P \in \mathcal{P}$ for each of the eight uncertainty sets, all eight configurations violate unichain and uniform ergodicity assumptions over all feasible kernels and stationary policies.

For KL and χ^2 uncertainty, we set δ to half the corresponding upper bound in Assumption 1, using the minimum positive transition probability of P. These radii preserve the nominal transition supports. For TV and Wasserstein uncertainty, we use $\delta = 0.1$ and allow transitions outside the nominal support. The Wasserstein order is $\ell = 2$, with ground metric $\rho(s, s') = \mathbb{1}\{s \neq s'\}$.

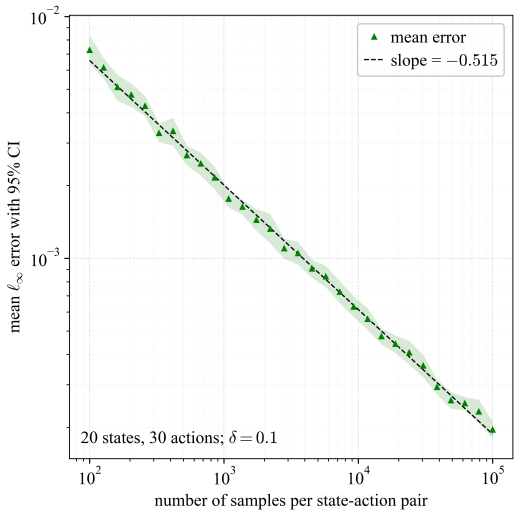
For each of 30 sample sizes n from 10^2 to 10^5, we independently draw n next-state samples from each $p_{s,a}$ and construct $\widehat{p}_{s,a}$ from empirical frequencies, without smoothing. We repeat this procedure over 20 independent trials. With $\gamma_n = 1 - n^{-1/2}$, we report the mean of $\|(1-\gamma_n)\widehat{v}_{\gamma_n}^* - g_\delta^* \mathbf{1}\|_\infty$



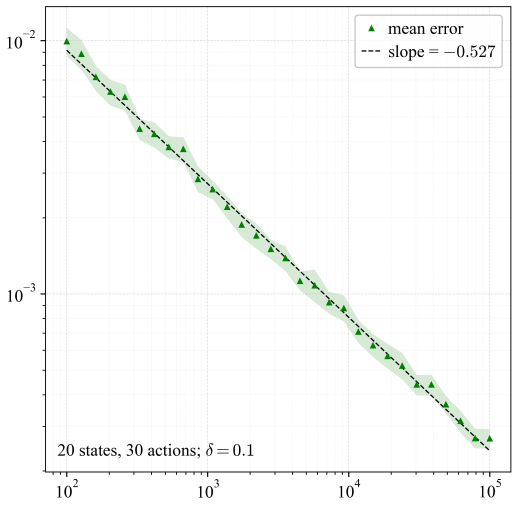
(a) SA-rectangular KL-divergence uncertainty.



(b) SA-rectangular χ^2-divergence uncertainty.



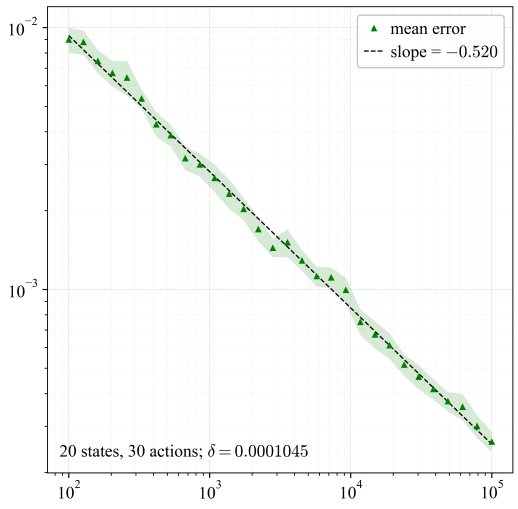
(c) SA-rectangular TV uncertainty.



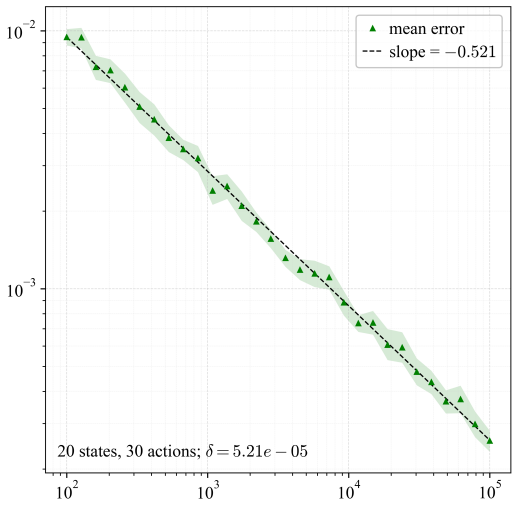
(d) SA-rectangular Wasserstein uncertainty.

Figure 3: Large-scale experiments under SA-rectangular uncertainty. Each panel plots the mean ℓ_∞ gain-estimation error against the number of samples per state-action pair. Shaded bands indicate 95% confidence intervals, and dashed lines show fitted slopes on logarithmic axes.

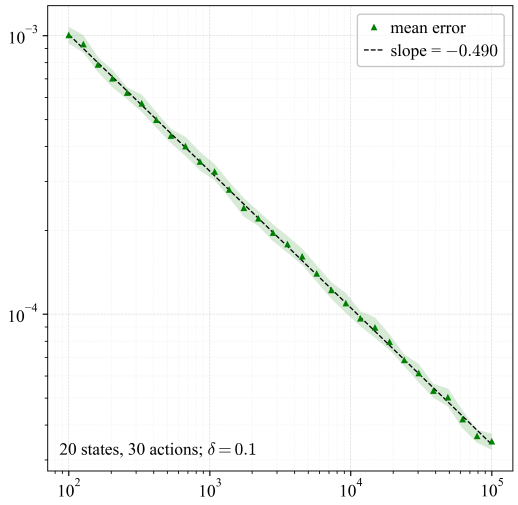
across trials, with percentile-bootstrap 95% confidence intervals. Figures 3 and 4 show the results. The fitted slopes, computed using all 30 sample sizes, range from -0.527 to -0.490, consistent with an error decay of approximately $n^{-1/2}$.



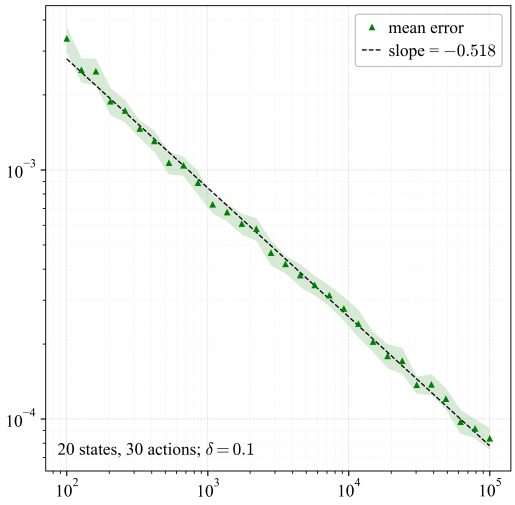
(a) S-rectangular KL-divergence uncertainty.



(b) S-rectangular χ^2-divergence uncertainty.



(c) S-rectangular TV uncertainty.



(d) S-rectangular Wasserstein uncertainty.

Figure 4: Large-scale experiments under S-rectangular uncertainty, with the same error measure and plotting conventions as Figure 3.