LIMITS OF ONLINE DISTRIBUTIONALLY ROBUST AVERAGE-REWARD REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We establish matching minimax horizon exponents for single-trajectory robust average-reward learning on tabular MDP families with bounded parameters. The learner observes transitions from an unknown nominal model P, while stationary policies are evaluated by their worst-case average reward over a rectangular f-divergence ball of radius δ. The deletion cost θ_f measures the minimum divergence needed to remove a transition, while the asymptotic slope $f'(\infty)$ determines whether new transitions can be created. These yield three exact radii $\delta_f^*(P) \le \delta_f^\dagger(P) \le \delta_f^{\mathrm{bell}}(P)$, characterizing, respectively, a common transition set guaranteeing weak communication, weak communication of every plausible model, and solvability of the robust Bellman equation with a state-independent gain for every bounded reward. For suitable fixed state and action counts and parameter bounds, minimax policy regret is $\widetilde{\Theta}(\sqrt{T})$ when transition creation is impossible and $\delta < \delta_f^{\mathrm{bell}}(P)$, even with nominal transient states. With creation and $0 < \delta < \theta_f(1)$, the rate is $\widetilde{\Theta}(T^{2/3})$ for communicating nominal MDPs, while linear regret can be unavoidable otherwise. At $\delta \ge \theta_f(1)$, every transition model is admissible and zero regret is achievable. A Bellman completion construction establishes the square-root upper bound beyond both communication thresholds.

1 INTRODUCTION

Robust MDPs optimize against the worst transition kernel in an ambiguity set (Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013), addressing model mismatch (Dulac-Arnold et al., 2021). Discounted robust RL has well-developed statistical guarantees (Panaganti et al., 2022; Wang et al., 2024b; Shi & Chi, 2024). Without discounting, however, the Bellman operator lacks the usual strict contraction in the supremum norm (Puterman, 1994). Moreover, long-run rewards depend on the recurrent classes induced by a policy and transition kernel, and on the probabilities of reaching them. Deleting or creating a transition can change this structure and make average reward discontinuous.

Recent work has established finite-sample guarantees for robust average-reward policy evaluation and control with access to a simulator that can sample any state–action pair (Roch et al., 2025; Chen et al., 2025; Roch et al., 2026; Xu et al., 2025; 2026; Yang et al., 2026). With a single trajectory, data collection depends on the policies executed by the learner. A policy's worst-case average reward may depend on transitions at states it never visits under the nominal kernel. Estimating its robust value may therefore require executing other policies. We study robust policy regret when an unknown nominal kernel P is observed through a single trajectory and rewards are known. Each announced stationary policy is evaluated against the ambiguity set

$$\mathcal{U}^P := \{Q : D_f(Q(\cdot \mid s,a) \,\|\, P(\cdot \mid s,a)) \le \delta \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A}\},$$

where D_f denotes the f-divergence. The optimal robust average reward starting from state s is

$$g^*_{\mathrm{rob},P}(s) := \sup_{\pi \in \Pi} \inf_{Q \in \mathcal{U}^P} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^\pi_Q \left[\sum_{t=0}^{T-1} r(S_t, A_t) \,\middle|\, S_0 = s \right],$$

where Π denotes the set of stationary randomized policies. Our regret measures the cumulative shortfall of the announced policies' robust average rewards relative to the optimum; Section 5 gives the formal definition. Our main contributions are:

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.

- We identify three sharp radii $\delta_f^*(P) \le \delta_f^\dagger(P) \le \delta_f^{\mathrm{bell}}(P)$ governing weak communication of the robust core (the transitions present in every ambiguity kernel), weak communication of every ambiguity kernel, and solvability of the robust Bellman equation with a common gain for every bounded reward.
- We establish matching minimax policy-regret rates, up to logarithmic factors, over families with bounded instance parameters. The rate is $\sqrt{T}$ when ambiguity cannot add transitions and $\delta < \delta_f^{\mathrm{bell}}(P)$. When ambiguity can add transitions, the rate is $T^{2/3}$ if the nominal MDP is communicating and the budget is below the level at which every transition distribution becomes feasible. Crossing $\delta_f^*(P)$ or $\delta_f^\dagger(P)$ does not change these exponents.
- We develop two algorithms that learn the transition support needed for robust control and attain these rates: one updates optimistic policies over successive episodes, and the other explores and then commits to a fixed robust policy.
- We characterize when learning is unnecessary and when sublinear regret is impossible. When the budget makes every transition distribution feasible, a policy that maximizes the known reward at each state has zero regret. For smaller budgets, allowing new transitions can force linear regret when the nominal MDP is weakly communicating but not communicating.

2 Related Work

Rectangular robust dynamic programming originates with Nilim & El Ghaoui (2005); Iyengar (2005). Learning guarantees depend on how transition data are obtained. Discounted robust RL has generative and offline guarantees (Panaganti & Kalathil, 2022; Wang et al., 2024b; Shi & Chi, 2024). With a single trajectory, Wang & Zou (2021) give finite-time bounds for contamination uncertainty under ergodicity and coverage conditions, while Liang et al. (2024) establish asymptotic convergence for f-divergence uncertainty.

In finite-horizon online learning, Lu et al. (2024) use a vanishing-value condition and He et al. (2025) impose visitation conditions. For KL and χ^2 uncertainty, ORVIT gives guarantees without these restrictions (Ghosh et al., 2026). Sun et al. (2025) study known transition prototypes and shrinking ambiguity sets, with regret measured under the true kernel. The average-reward setting studied here also requires accounting for recurrent classes and the existence of a common optimal gain.

Robust average-reward work establishes optimality and convergence results (Tewari & Bartlett, 2007; Wang et al., 2024c; Grand-Clément et al., 2026), as well as the generative-sampling guarantees discussed above. Earlier continuing online work combines stochastic and adversarial state–action pairs and measures regret through realized rewards (Lim et al., 2016). Wang & Si (2026) connect regret guarantees to robust optimality for a prescribed nonrectangular ambiguity set. We learn an unknown ambiguity-set center from one nominal trajectory and evaluate each announced stationary policy by its robust average reward.

For Bellman solvability, Wang & Si (2025) give constant-gain theory under communication conditions. We compute the exact budget threshold for solvability for every bounded reward and distinguish it from the two communication thresholds. The proof specializes nonlinear ergodicity theory (Akian et al., 2015a;b) by converting its common-gain criterion into the explicit radius $\delta_f^{\mathrm{bell}}(P)$. Our $T^{2/3}$ lower bound reflects the cost of choosing informative policies with lower robust value, paralleling the revealing-action tradeoff in stochastic partial monitoring (Bartók et al., 2011). Appendix A gives the full comparison with robust and nominal learning results.

3 Preliminaries

An MDP is a tuple $(\mathcal{S}, \mathcal{A}, r, Q)$ with finite nonempty state and action spaces $\mathcal{S}$ and $\mathcal{A}$, a known deterministic reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and a transition kernel $Q : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. Write $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$. We use Q for a generic kernel and P for the unknown nominal kernel that generates the learner's observations. The ambiguity sets are centered at the transition distributions $P(\cdot \mid s, a)$. Assuming known rewards lets us isolate the challenges arising from transition uncertainty.

The stationary randomized and deterministic policy classes are $\Pi := \Delta(\mathcal{A})^{\mathcal{S}}$ and $\Pi_{\mathrm{d}} := \mathcal{A}^{\mathcal{S}}$, respectively, with each deterministic policy $\pi \in \Pi_{\mathrm{d}}$ embedded in Π via $\pi(\cdot \mid s) := \mathbf{e}_{\pi(s)}$. For a

kernel Q and policy $\pi \in \Pi$, write $Q_\pi(s' \mid s) := \sum_a \pi(a \mid s) Q(s' \mid s, a)$ and $Q_{s,a} := Q(\cdot \mid s, a)$. Define the average reward by

$$g_Q^\pi(s) := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_Q^\pi \left[\sum_{t=0}^{T-1} r(S_t, A_t) \mid S_0 = s \right].$$

Set $g_Q^*(s) := \sup_{\pi\in\Pi} g_Q^\pi(s)$ and, for $u \in \mathbb{R}^{\mathcal{S}}$, let $\mathrm{sp}(u) := \max_s u(s) - \min_s u(s)$.

Let $\mathcal{F}$ be the class of proper, lower semicontinuous convex functions $f : [0,\infty) \to \mathbb{R} \cup \{+\infty\}$ that are finite on $(0,\infty)$, satisfy $f(1) = 0$ and $tf(x_0)+(1-t)f(x_1) > 0$ whenever $x_0, x_1 \in [0,\infty)$, $x_0 \neq x_1$, $t \in (0,1)$, and $tx_0 + (1-t)x_1 = 1$. This strict Jensen condition at 1 includes the TV generator $f(x) = |x-1|/2$. We use the boundary values $f(0) := \lim_{x\downarrow 0} f(x)$, and $f'(\infty) := \lim_{x\to\infty} f(x)/x$, where $f'(\infty)$ is the asymptotic slope. For $p, q \in \Delta(\mathcal{S})$, with the convention $0 \cdot (+\infty) = 0$, define

$$D_f(p\|q) := \sum_{i:q_i>0} q_i f(p_i/q_i) + f'(\infty) \sum_{i:q_i=0} p_i.$$

These boundary values restrict how the adversary can change the nominal support. Deleting a transition with nominal probability $q_i > 0$ contributes $q_i f(0)$ to the divergence, while assigning probability $p_i > 0$ to a nominally absent successor contributes $p_i f'(\infty)$. Thus, at every finite radius, $f(0) = +\infty$ forbids deletion, while $f'(\infty) = +\infty$ forbids creation. We define the non-collapsing and non-expanding classes, respectively, by $\mathcal{F}_{\mathrm{nc}} := \{f \in \mathcal{F} : f(0) = +\infty\}$ and $\mathcal{F}_{\mathrm{ne}} := \{f \in \mathcal{F} : f'(\infty) = +\infty\}$. Deleting transitions also requires redistributing their probability mass. The minimum total divergence, including this redistribution, is $\theta_f(x)$ when the deleted successors carry total nominal probability $x \in (0,1)$; see Section 4. For $0 < \delta < \infty$, let

$$\mathcal{U}^q(\delta) := \{p \in \Delta(\mathcal{S}) : D_f(p\|q) \le \delta\}, \qquad \mathcal{U}^P(\delta) := \{\widetilde{Q} : \widetilde{Q}_{s,a} \in \mathcal{U}^{P_{s,a}}(\delta)\ \forall (s,a)\},$$

abbreviated $\mathcal{U}^P$ when δ is fixed. The product form allows each state–action row to vary independently; this is state–action rectangularity. For $\pi \in \Pi$, define its robust gain and the robust optimal gain by

$$g_{\mathrm{rob},P}^\pi(s) := \inf_{\widetilde{Q}\in\mathcal{U}^P} g_{\widetilde{Q}}^\pi(s), \qquad g_{\mathrm{rob},P}^*(s) := \sup_{\pi\in\Pi} g_{\mathrm{rob},P}^\pi(s).$$

We use the same notation with any center Q in place of P. For $q \in \Delta(\mathcal{S})$ and $u \in \mathbb{R}^{\mathcal{S}}$, set $\sigma_\delta^f(q,u) := \min\{p^\top u : p \in \mathcal{U}^q(\delta)\}$. The minimum is attained because $\mathcal{U}^q(\delta)$ is nonempty and compact. For $\gamma \in [0,1]$, define

$$\mathcal{L}_{Q,\gamma}^{\mathrm{rob}} u(s) := \max_{a\in\mathcal{A}} \{r(s,a) + \gamma \sigma_\delta^f(Q_{s,a}, u)\}, \qquad \mathcal{L}_Q^{\mathrm{rob}} := \mathcal{L}_{Q,1}^{\mathrm{rob}}.$$

A pair $(g,h) \in \mathbb{R} \times \mathbb{R}^{\mathcal{S}}$ is an *additive eigenpair* of an operator $\mathcal{T} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$, written $(g,h) \in \mathrm{Eig}(\mathcal{T})$, when $g\mathbf{1} + h = \mathcal{T}h$, where $\mathbf{1}$ is the all-ones vector. A deterministic policy $\pi \in \Pi_{\mathrm{d}}$ is *greedy* for h at Q, written $\pi \in \mathrm{Gr}_Q(h)$, when $\pi(s) \in \arg\max_{a\in\mathcal{A}} \{r(s,a) + \sigma_\delta^f(Q_{s,a}, h)\}$, for all $s \in \mathcal{S}$.

If $(g,h) \in \mathrm{Eig}(\mathcal{L}_Q^{\mathrm{rob}})$, then every $\pi \in \mathrm{Gr}_Q(h)$ satisfies $g_{\mathrm{rob},Q}^\pi(s) = g_{\mathrm{rob},Q}^*(s) = g$ for every state s. Weak communication of the nominal kernel Q guarantees an additive eigenpair for the nominal Bellman operator, but need not do so for $\mathcal{L}_Q^{\mathrm{rob}}$. Section 4 identifies the exact ambiguity radius below which the robust Bellman equation admits an additive eigenpair for every bounded reward.

4 Three thresholds for robust control

For a fixed nominal kernel, increasing the adversarial budget enlarges the set of feasible transition kernels. This can change which states are reachable from one another under these kernels and whether the optimal robust average reward depends on the initial state. In this section, we identify three thresholds that describe these changes.

A transition graph G assigns to every (s,a) a nonempty set $G(s,a) \subseteq \mathcal{S}$; write $\mathfrak{G}$ for the collection of these graphs. We identify G with its edge set $\{(s,a,s') : s' \in G(s,a)\}$, so $H \subseteq G$ means $H(s,a) \subseteq G(s,a)$ for every (s,a). For an edge $e = (s,a,s')$, write $Q(e) := Q(s' \mid s,a)$.

A nonempty set $I \subseteq \mathcal{S}$ is *closed* in G if $G(s,a) \subseteq I$ for every $s \in I$ and $a \in \mathcal{A}$. A nonempty set $J \subseteq \mathcal{S}$ is *closable* if, for every $s \in J$, some action a satisfies $G(s,a) \subseteq J$. Thus a closed set cannot be left under any policy, whereas a closable set can be kept invariant by a suitable stationary deterministic policy. We call G *weakly communicating* if there are no disjoint nonempty sets I, J with I closed and J closable, and write $\mathcal{M}$ for this class. Every transition graph is the support graph of some kernel; graphs with an empty row are excluded from $\mathcal{M}$.

For $\pi \in \Pi_{\mathrm{d}}$ write $G_\pi(s) := G(s, \pi(s))$. The recurrent classes $\mathrm{Rec}(G_\pi)$ are the minimal nonempty closed sets of G_π, and π is *unichain in* G when $|\mathrm{Rec}(G_\pi)| = 1$. Splitting $\mathcal{S}$ into the states recurrent under some stationary deterministic policy and the rest gives

$$C(G) := \bigcup_{\pi \in \Pi_{\mathrm{d}}} \bigcup_{R \in \mathrm{Rec}(G_\pi)} R, \qquad T(G) := \mathcal{S} \setminus C(G).$$

Thus $\mathcal{S} = C(G) \sqcup T(G)$, where every state in $T(G)$ is transient under every stationary deterministic policy. When $G \in \mathcal{M}$, the set $C(G)$ is closed under every action, and each of its states is reachable from every other under a suitable stationary deterministic policy (Puterman, 1994). For any kernel Q, define its support graph by $\mathcal{G}(Q)(s,a) := \mathrm{supp}(Q_{s,a})$ and set $C(Q) := C(\mathcal{G}(Q))$. We call Q communicating when $\mathcal{G}(Q) \in \mathcal{M}$ and $C(Q) = \mathcal{S}$. For $\pi \in \Pi_{\mathrm{d}}$ unichain in $\mathcal{G}(Q)$, let μ_Q^π denote the stationary distribution of Q_π, satisfying $(\mu_Q^\pi)^\top Q_\pi = (\mu_Q^\pi)^\top$.

All learning, regret, and minimax results assume $\mathcal{G}(P) \in \mathcal{M}$, except Theorem 5(iii). The structural results below apply to arbitrary nominal kernels. We impose $C(P) = \mathcal{S}$ only where stated.

The *robust core* of Q is the intersection of the support graphs of all ambiguity kernels:

$$\mathcal{G}_\wedge(Q) := \bigcap_{\widetilde{Q} \in \mathcal{U}^Q} \mathcal{G}(\widetilde{Q}).$$

It contains exactly the transitions present in every ambiguity kernel. Some rows may be empty, in which case $\mathcal{G}_\wedge(Q) \notin \mathfrak{G}$ and $\mathcal{G}_\wedge(Q) \notin \mathcal{M}$. Define $\theta_f(x) := xf(0) + (1-x)f(\frac{1}{1-x})$ for $x \in [0,1)$, with $\theta_f(0) = 0$ under the convention $0 \cdot (+\infty) = 0$, and $\theta_f(1) := \lim_{x \uparrow 1} \theta_f(x) = f(0) + f'(\infty)$. Lemma 2 shows that deleting transitions with total nominal probability $x \in (0,1)$ from a single row requires a minimum divergence of $\theta_f(x)$, including the cost of redistributing their probability. Lemma 3 shows that a transition absent from the nominal model can be assigned positive probability at any radius $\delta > 0$ if and only if $f'(\infty) < +\infty$.

Let $q_\star(f, \delta)$ denote the largest nominal probability mass whose deletion cost is at most δ, i.e.,

$$q_\star = q_\star(f,\delta) := \sup\{x \in [0,1] : \theta_f(x) \le \delta\}. \tag{1}$$

Since θ_f is nondecreasing and lower semicontinuous on $[0,1]$, $\theta_f(x) \le \delta$ exactly when $x \le q_\star$. In particular, a nominal edge e can be deleted if and only if $P(e) \le q_\star$. Moreover, $q_\star = 0$ exactly when $f \in \mathcal{F}_{\mathrm{nc}}$, and $q_\star < 1$ exactly when $\delta < \theta_f(1)$. These properties are proved in Appendix C. Table 1 in the appendix gives θ_f, $q_\star$, and the deletion/creation class for standard divergence generators. The comparison $P(e) \le q_\star$ is how the deletion threshold enters the sample-complexity results below.

The edge-deletion costs determine the robust core threshold (with $\max \varnothing := 0$):

$$\delta_f^*(P) := \max\Big\{\min_{e \in H} \theta_f(P(e)) : H \subseteq \mathcal{G}(P),\ H \in \mathcal{M}\Big\}.$$

It suffices to retain all nominal edges with deletion cost at least a chosen threshold. The next lemma justifies this reduction using the fact that adding edges preserves weak communication.

Lemma 1. *If $H \subseteq H'$ and $H \in \mathcal{M}$, then $H' \in \mathcal{M}$. Consequently, writing $H_{\ge\theta} := \{e \in \mathcal{G}(P) : \theta_f(P(e)) \ge \theta\}$, we have $\delta_f^*(P) = \max\Big\{\theta \in \{\theta_f(P(e)) : e \in \mathcal{G}(P)\} : H_{\ge\theta} \in \mathcal{M}\Big\}$.*

Theorem 1. *For every P, $f \in \mathcal{F}$, and $\delta > 0$, $\mathcal{G}_\wedge(P) = \{e \in \mathcal{G}(P) : \theta_f(P(e)) > \delta\}$, and*

$$\mathcal{G}_\wedge(P) \in \mathcal{M} \iff \delta < \delta_f^*(P) \implies \mathcal{U}^P(\delta) \subseteq \{Q : \mathcal{G}(Q) \in \mathcal{M}\}.$$

Moreover, quantifying over all instances,

$$f \in \mathcal{F}_{\mathrm{nc}} \iff \big[\mathcal{G}_\wedge(Q) = \mathcal{G}(Q)\ \ \forall Q,\ \forall \delta \in (0,\infty)\big], \tag{2}$$

$$f \in \mathcal{F}_{\mathrm{ne}} \iff \big[\mathcal{G}(\widetilde{Q}) \subseteq \mathcal{G}(Q)\ \ \forall Q,\ \forall \delta \in (0,\infty),\ \forall \widetilde{Q} \in \mathcal{U}^Q(\delta)\big], \tag{3}$$

so $\mathcal{F}_{\mathrm{nc}}$ is exactly the no-deletion class and $\mathcal{F}_{\mathrm{ne}}$ the no-creation class. These characterizations require the properties to hold for all kernels and radii.

The robust core $\mathcal{G}_\wedge(P)$ contains the transitions that cannot be deleted within adversarial budget δ. It is weakly communicating exactly when $\delta < \delta_f^*(P)$, ensuring weak communication of every ambiguity kernel. In this regime, write $C_\wedge(P) := C(\mathcal{G}_\wedge(P)) \subseteq C(P)$; every state outside $C_\wedge(P)$ is transient under every stationary deterministic policy in $\mathcal{G}_\wedge(P)$. At or above the threshold, every ambiguity kernel may still be weakly communicating: different kernels may delete different edges, and those deletions need not be jointly feasible within the same budget.

By Lemma 1, $\delta_f^*(P)$ can be computed using $O(S^3 A \log(SA))$ graph operations: sort the at most S^2A edge costs, binary-search the nested threshold graphs, and test weak communication using the fixed-point procedure in the proof.

An ambiguity kernel fails to be weakly communicating precisely when it has two disjoint nonempty sets: one closed under every action, and another that a deterministic policy can keep closed. To determine the divergence budgets needed to construct these sets, define, for nonempty $B \subseteq \mathcal{S}$,

$$\rho^P_{f,\mathrm{all}}(B) := \max_{s\in B,\, a\in\mathcal{A}} \theta_f\big(P(B^c \mid s,a)\big), \qquad \rho^P_{f,\mathrm{pol}}(B) := \max_{s\in B} \min_{a\in\mathcal{A}} \theta_f\big(P(B^c \mid s,a)\big).$$

The quantity $\rho^P_{f,\mathrm{all}}(B)$ is the smallest rowwise radius that allows B to be closed under every action in some ambiguity kernel. The quantity $\rho^P_{f,\mathrm{pol}}(B)$ is the corresponding radius for keeping B closed under some deterministic policy. We call B δ-*closed* when $\rho^P_{f,\mathrm{all}}(B) \le \delta$, and δ-*closable* when $\rho^P_{f,\mathrm{pol}}(B) \le \delta$. Since $\rho^P_{f,\mathrm{pol}}(B) \le \rho^P_{f,\mathrm{all}}(B)$, every δ-closed set is also δ-closable.

For disjoint sets I, J, the required row changes can be made independently. The smallest radius admitting such a pair is therefore

$$\delta_f^\dagger(P) := \min_{I,J\neq\varnothing,\, I\cap J=\varnothing} \max\left\{\rho^P_{f,\mathrm{all}}(I), \rho^P_{f,\mathrm{pol}}(J)\right\}.$$

Throughout, minima over empty families equal $+\infty$.

Theorem 2. *For every P, $f \in \mathcal{F}$ and $\delta > 0$, $\mathcal{U}^P(\delta) \not\subseteq \{Q : \mathcal{G}(Q) \in \mathcal{M}\} \iff \delta \ge \delta_f^\dagger(P)$.*

Weak communication of every ambiguity kernel guarantees that, for every bounded reward, the robust Bellman equation $g\mathbf{1} + h = \mathcal{L}_P^{\mathrm{rob}} h$ admits a solution $(g,h) \in \mathbb{R}\times\mathbb{R}^{\mathcal{S}}$. However, this condition is not necessary. The equation has no such solution for some bounded reward exactly when there are two disjoint nonempty sets I, J: the adversary can keep the process in I regardless of the controller's actions, while the controller can keep it in J regardless of the adversary's feasible transition choices. For $f \in \mathcal{F}_{\mathrm{ne}}$, the controller can keep the process in J under every ambiguity kernel exactly when J is closable in $\mathcal{G}(P)$. To keep the process in I regardless of the controller's action, the adversary needs a feasible row supported on I for every $s \in I$ and $a \in \mathcal{A}$. This is possible exactly when $\delta \ge \rho^P_{f,\mathrm{all}}(I)$. For $f \notin \mathcal{F}_{\mathrm{ne}}$, the controller cannot keep the process in any nonempty proper set against all feasible transition rows, since the adversary can always assign positive probability outside it. These observations give the Bellman radius: set $\delta_f^{\mathrm{bell}}(P) := +\infty$ when $f \notin \mathcal{F}_{\mathrm{ne}}$. Otherwise, define

$$\delta_f^{\mathrm{bell}}(P) := \min_{I,J\neq\varnothing,\, I\cap J=\varnothing,\, J \text{ closable in } \mathcal{G}(P)} \rho^P_{f,\mathrm{all}}(I). \tag{4}$$

Theorem 3. *Fix P, $f \in \mathcal{F}$, and $\delta > 0$. The following are equivalent:*

(i) $\delta < \delta_f^{\mathrm{bell}}(P)$*;*

(ii) every fixed point of the recession operator, $[\widehat{\mathcal{L}}_P u](s) := \max_{a\in\mathcal{A}} \sigma_\delta^f(P_{s,a}, u)$*, is constant;*

(iii) for every reward $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$*,* $\mathrm{Eig}(\mathcal{L}_P^{\mathrm{rob}})$ *is nonempty and all its eigenpairs share the same gain* $g^{\mathrm{rob}}(P)$*. For every* $(g^{\mathrm{rob}}(P), h) \in \mathrm{Eig}(\mathcal{L}_P^{\mathrm{rob}})$ *and every* $\pi^* \in \mathrm{Gr}_P(h)$*,*

$$g^{\pi^*}_{\mathrm{rob},P}(s) = g^*_{\mathrm{rob},P}(s) = g^{\mathrm{rob}}(P), \qquad s \in \mathcal{S}.$$

Under these equivalent conditions, fix a bounded reward r and a reference state $s_\circ \in \mathcal{S}$. For each $\gamma \in (0,1)$, let v_γ be the unique fixed point of $\mathcal{L}^{\mathrm{rob}}_{P,\gamma}$. Then $(1-\gamma)v_\gamma \to g^{\mathrm{rob}}(P)\mathbf{1}$ as $\gamma \uparrow 1$, the normalized values $\{v_\gamma - v_\gamma(s_\circ)\mathbf{1} : \gamma \in (0,1)\}$ are uniformly bounded, and every accumulation point as $\gamma \uparrow 1$ belongs to the normalized bias set

$$\mathcal{H}(P) := \{h \in \mathbb{R}^{\mathcal{S}} : h(s_\circ) = 0,\ (g^{\mathrm{rob}}(P), h) \in \mathrm{Eig}(\mathcal{L}_P^{\mathrm{rob}})\}.$$

Conversely, if $\delta \geq \delta_f^{\mathrm{bell}}(P)$, there is a reward $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$ for which the optimal robust average reward depends on the initial state.

In particular, if $f'(\infty) < +\infty$, then $\delta_f^{\mathrm{bell}}(P) = +\infty$ for every P. Thus, for every positive finite radius and every bounded reward, the robust Bellman equation has a solution with one gain for all states, even if some ambiguity kernels are not weakly communicating.

Corollary 1. *For every $f \in \mathcal{F}$, we have $\delta_f^*(P) \leq \delta_f^\dagger(P) \leq \delta_f^{\mathrm{bell}}(P)$. Within the no-creation class $f \in \mathcal{F}_{\mathrm{ne}}$, either inequality can be strict; for KL, both can be strict simultaneously. Such strict separations need not occur for every generator in the class. Even within $\mathcal{F}_{\mathrm{ne}}$, $\delta_f^\dagger(P) < \delta_f^{\mathrm{bell}}(P) = +\infty$ can occur. For $f \notin \mathcal{F}_{\mathrm{ne}}$, $\delta_f^{\mathrm{bell}}(P) = +\infty$ for every P, while $\delta_f^\dagger(P)$ may be finite. When $f \in \mathcal{F}_{\mathrm{nc}} \cap \mathcal{F}_{\mathrm{ne}}$ and $\mathcal{G}(P)$ is weakly communicating, all three radii equal $+\infty$.*

Thus δ_f^* is the threshold for a common weakly communicating core, $\delta_f^\dagger$ is the threshold for weak communication of every ambiguity kernel, and δ_f^{bell} is the threshold below which the robust Bellman equation has a solution for every bounded reward. Examples establishing strict inequalities appear in Proposition 2 and the proof of Corollary 1.

Policies greedy with respect to a robust bias in $\mathcal{H}(P)$ need not be unichain, even without ambiguity (Appendix Proposition 3). Thus the existence of a scalar robust optimal gain does not justify assuming a unique stationary distribution under every greedy policy.

5 Topology-adaptive online learning

Building on these structural results, we develop two algorithms for learning robust policies from a single trajectory under the unknown nominal kernel. We first define policy regret .

Suppose the optimal robust average reward has the same value $g^{\mathrm{rob}}(P)$ from every initial state. For a learner that announces a stationary policy π_t at each round t, define

$$\mathrm{Reg}_{\mathrm{pol}}(T) := \sum_{t=0}^{T-1} \Big[g^{\mathrm{rob}}(P) - g^{\pi_t}_{\mathrm{rob},P}(S_t) \Big]. \tag{5}$$

Let $H_t = (S_0, A_0, \ldots, A_{t-1}, S_t)$ and $\mathcal{H}_t := \mathcal{S} \times (\mathcal{A} \times \mathcal{S})^t$. A learner is a sequence of measurable maps $\mathfrak{A}_t : [0,1]^{\mathcal{S}\times\mathcal{A}} \times \mathcal{H}_t \times [0,1] \to \Pi$. Given the known reward r, the interaction history H_t, and a private seed $U \sim \mathrm{Unif}[0,1]$ independent of the environment, it announces a stationary policy $\pi_t = \mathfrak{A}_t(r, H_t, U)$ and draws $A_t \sim \pi_t(\cdot \mid S_t)$. We write $\mathfrak{A}_{\mathrm{pol}}$ for the class of all such learners.

At round t, the learner takes one step under the nominal kernel using π_t. Its regret compares the optimal robust gain with the robust average reward of following π_t indefinitely from S_t. This allows policies with multiple recurrent classes and evaluates the announced policies, rather than the accumulated rewards. Actions follow the announced policy. The learner may change this policy at every round, including choosing deterministic policies to collect information.

We consider two algorithms that begin by collecting transition data through uniform exploration. Let $R(P) := C(P)$ when $f \in \mathcal{F}_{\mathrm{ne}}$ and $R(P) := \mathcal{S}$ otherwise, and let P_R denote the restriction of P to $R(P)$. Algorithm 3 explores for $\lceil \sqrt{T} \rceil$ rounds and then updates an optimistic policy over successive episodes. Algorithm 1 explores for $\lceil T^{2/3} \rceil$ rounds and then follows a fixed robust policy for the remaining rounds. Under the standing assumption $\mathcal{G}(P) \in \mathcal{M}$, the no-creation guarantees do not require $C(P) = \mathcal{S}$. For $f \notin \mathcal{F}_{\mathrm{ne}}$, the regret guarantee in Theorem 4(ii) additionally assumes $C(P) = \mathcal{S}$. Appendix D gives the pseudocode and proves Theorem 4.

Theorem 4. *Fix $\alpha \in (0,1)$, write $d_R := |R(P)|$, and set $m_T^{\mathrm{BC}} := \lceil \sqrt{T} \rceil$ and $m_T := \lceil T^{2/3} \rceil$. For each fixed instance, the bounds below hold for horizons above their respective finite thresholds. These thresholds and the constants and functions used below are specified in Appendices B and D.*

(i) Suppose $f \in \mathcal{F}_{\mathrm{ne}}$ and $\delta < \delta_f^{\mathrm{bell}}(P)$. Write $\mathfrak{C}_{\mathrm{ne}}(P_R) := H_{\mathrm{loc}}(P_R) K_{\mathrm{face}}(P_R) C_{\mathrm{occ}}(P_R)$. Then Algorithm 3 satisfies

$$\mathbb{E}\,\mathrm{Reg}_{\mathrm{pol}}(T) \leq m_T^{\mathrm{BC}} + \widetilde{O}\Big(\mathfrak{C}_{\mathrm{ne}}(P_R) \sqrt{d_R A T [S + \log(SAT/\alpha)]} \Big) + \widetilde{O}\big(\mathfrak{C}_{\mathrm{ne}}(P_R) M_{\mathrm{mix}}(P_R) d_R^2 A \big) + \alpha T.$$

Moreover, $C_{\mathrm{occ}}(P_R) \le 1 + D_{\mathrm{cmp}}(P_R)$, $M_{\mathrm{mix}}(P_R) \le 2D_{\mathrm{cmp}}(P_R)$, *and* $K_{\mathrm{face}}(P_R) \le 2/p_{\min,R}$, *where* $p_{\min,R}$ *is the smallest positive entry of* P_R.

(ii) Suppose $\mathcal{G}(P) \in \mathcal{M}$, $\delta < \delta_f^{\mathrm{bell}}(P)$, *and either* $f \in \mathcal{F}_{\mathrm{ne}}$ *or* $C(P) = \mathcal{S}$. *Set* $\eta_T := \beta_{\alpha/3}(n_T(P,\alpha))$. *Then Algorithm 1 satisfies, with probability at least* $1-\alpha$,

$$\mathrm{Reg}_{\mathrm{pol}}(T) \le m_T + 2(T - m_T)H_{\mathrm{loc}}(P_R)\varpi_{f,\delta}^{P_R}(\eta_T) = o(T).$$

Here $n_T(P,\alpha)$ *is a high-probability lower bound on exploration samples per row of* P_R, $\beta_\alpha(\cdot)$ *bounds the* ℓ_1 *estimation error, and* $\varpi_{f,\delta}^{P_R}(\cdot)$ *bounds its local effect on worst-case expectations for vectors of span at most one; see Appendix B.*

Under the additional support and accuracy conditions of Appendix Theorem 9, the same probability guarantee gives

$$\mathrm{Reg}_{\mathrm{pol}}(T) \le \widetilde{O}\Big(\Big[1 + H_{\mathrm{loc}}(P_R)K_{\mathrm{face}}(P_R)\sqrt{D_0(P)d_R A[S + \log(SAT/\alpha)]}\Big]\, T^{2/3}\Big),$$

where $D_0(P)-1$ *is the maximum expected time under uniform exploration to visit a specified pair in* $C(P) \times \mathcal{A}$, *over all target pairs and initial states.*

All bounds hold uniformly in the initial state. For each fixed instance, $\alpha = T^{-2}$ *gives expected regret* $\widetilde{O}(\sqrt{T})$ *in (i) and* $\widetilde{O}(T^{2/3})$ *in (ii).*

Identifying transitions that the adversary cannot delete requires testing whether their nominal probabilities exceed $q_\star$. When $q_\star = 0$, observing a transition certifies that it belongs to the robust core $\mathcal{G}_\wedge(P)$. When $0 < q_\star < 1$, however, the sample requirement depends on the separation between the transition probability and $q_\star$. Appendix Theorem 7 gives sharp sample bounds for these tests.

The regret guarantees do not imply polynomial running time. Evaluating $\delta_f^\dagger$ and δ_f^{bell} may require checking up to 3^S assignments to I, J, and their complement. The implementation of BellReady in Appendix D may also enumerate disjoint set pairs. We establish no polynomial-time algorithm for this step or for optimistic robust planning over the product confidence set.

6 Minimax rates

We complement the preceding regret guarantees with matching lower bounds in T. Together, these bounds characterize the minimax rates across the three structural regimes, up to logarithmic factors. The three regimes are

$$\delta < \delta_f^*(P) \quad (\mathrm{core}), \qquad \delta_f^*(P) \le \delta < \delta_f^\dagger(P) \quad (\mathrm{all}), \qquad \delta_f^\dagger(P) \le \delta < \delta_f^{\mathrm{bell}}(P) \quad (\mathrm{bell}).$$

The labels core, all, and bell denote, respectively, a weakly communicating common core, weak communication of every ambiguity kernel, and Bellman solvability for every bounded reward despite some ambiguity kernel not being weakly communicating. Set

$$\Xi_f := \begin{cases} \{\mathrm{core}\}, & f \in \mathcal{F}_{\mathrm{nc}}, \\ \{\mathrm{core}, \mathrm{all}, \mathrm{bell}\}, & f \notin \mathcal{F}_{\mathrm{nc}}, \end{cases} \qquad \alpha_f := \begin{cases} 1/2, & f \in \mathcal{F}_{\mathrm{ne}}, \\ 2/3, & f \notin \mathcal{F}_{\mathrm{ne}}. \end{cases}$$

To obtain uniform bounds, we control the instance constants from the preceding section. We define a family for each choice of state and action spaces and each $B \ge 1$. Define the structural margins

$$\Delta_\theta(P) := \min_{\substack{s \in C(P),\, a \in \mathcal{A} \\ s' \in \mathcal{G}_\wedge(P)(s,a)}} [\theta_f(P(s' \mid s,a)) - \delta], \qquad \Delta_{\mathrm{bell}}(P_R) := \min\{1, p_{\min,R}, m_{\mathrm{bell}}(P_R)\}.$$

These margins quantify stability under estimation error: Δ_θ measures how far unavoidable transitions are from becoming deletable, while Δ_{bell} controls how much transition probabilities can change without losing Bellman solvability, provided their support is unchanged. Set $m_{\mathrm{bell}}(P_R) := +\infty$ when $f \notin \mathcal{F}_{\mathrm{ne}}$; otherwise, define

$$m_{\mathrm{bell}}(P_R) := \min_{I,J \subseteq R(P),\, I,J \neq \varnothing,\, I \cap J = \varnothing,\, J \text{ closable in } \mathcal{G}(P_R)} \max_{s \in I,\, a \in \mathcal{A}} [P_R(R(P) \setminus I \mid s,a) - q_\star].$$

We use $\min \varnothing = +\infty$. When $f \in \mathcal{F}_{\mathrm{nc}}$, $\Delta_\theta(P) = +\infty$.

For a known reward r, the family $\mathfrak{P}_f^\xi(B; r)$ consists of kernels in regime ξ satisfying $\mathcal{G}(P) \in \mathcal{M}$, $D_0(P), H_{\mathrm{loc}}(P_R; r) \le B$, and $p_{\min,R}, \Delta_\theta(P) \ge B^{-1}$. For $\xi \in \{\mathrm{all}, \mathrm{bell}\}$, also require $\Delta_{\mathrm{bell}}(P_R) \ge B^{-1}$. For $f \in \mathcal{F}_{\mathrm{ne}}$, also require $D_{\mathrm{cmp}}(P_R) \le B$; for $f \notin \mathcal{F}_{\mathrm{ne}}$, require $C(P) = \mathcal{S}$. The minimax expected policy regret is

$$\mathfrak{R}_f^\xi(T; B) := \inf_{\mathfrak{A} \in \mathfrak{A}_{\mathrm{pol}}} \sup_{r \in [0,1]^{\mathcal{S}\times\mathcal{A}},\, s_0 \in \mathcal{S},\, P \in \mathfrak{P}_f^\xi(B;r)} \mathbb{E}_{P,r,s_0}^{\mathfrak{A}} \operatorname{Reg}_{\mathrm{pol}}(T).$$

Appendix D gives the full family definitions and bounds on the auxiliary constants. The rates below describe dependence on T with S, A, and B held fixed. The minimum horizon required for the regret bounds may grow without bound as B increases.

Corollary 2. *Choose $\alpha = 1/T$. Uniformly over the families defined above, for all sufficiently large T, Algorithms 3 and 1 satisfy, respectively,*

$$\mathbb{E} \operatorname{Reg}_{\mathrm{pol}}(T) \le \begin{cases} \widetilde{O}_{f,\delta}\Big(B^3\sqrt{SAT[S + \log(SAT)]} + B^4 S^2 A\Big), & f \in \mathcal{F}_{\mathrm{ne}}, \\ \widetilde{O}_{f,\delta}\big(\mathrm{poly}(B,S,A)T^{2/3}\big), & f \notin \mathcal{F}_{\mathrm{ne}},\ \delta < \theta_f(1). \end{cases}$$

For fixed B, S, A, the rates are $\widetilde{O}(\sqrt{T})$ and $\widetilde{O}(T^{2/3})$, respectively.

When $\delta \ge \theta_f(1)$, the adversary can choose any transition distribution at each state–action pair; we call this saturation. Lemma 20 then gives, for every nominal kernel P and stationary policy π,

$$\begin{aligned} \mathcal{U}^q(\delta) &= \Delta(\mathcal{S}), & \sigma_\delta^f(q,u) &= \min_x u(x), \\ g_{\mathrm{rob},P}^\pi(s) &= \min_x \sum_a \pi(a \mid x) r(x,a), & g^{\mathrm{rob}}(P) &= \min_x \max_a r(x,a). \end{aligned} \tag{6}$$

Theorem 5. *Fix $f \in \mathcal{F}$ and $\delta > 0$.*

(i) Matching regret rates: *Suppose $\delta < \theta_f(1)$. For every nonempty regime indexed by $\xi \in \Xi_f$, there are finite state and action dimensions S_ξ, A_ξ and constants $B_{0,\xi}, c_{f,\delta,\xi} > 0$ such that, for every $B \ge B_{0,\xi}$, there is a finite $T_{0,\xi}(B, f, \delta)$ for which the minimax regret at these dimensions satisfies*

$$c_{f,\delta,\xi} T^{\alpha_f} \le \mathfrak{R}_f^\xi(T; B) \le \widetilde{O}_{B,f,\delta,S_\xi,A_\xi}\big(T^{\alpha_f}\big), \qquad T \ge T_{0,\xi}. \tag{7}$$

For $f \in \mathcal{F}_{\mathrm{ne}}$, the exponent is $1/2$, without requiring $C(P) = \mathcal{S}$. For $f \notin \mathcal{F}_{\mathrm{ne}}$ and $C(P) = \mathcal{S}$, the exponent is $2/3$. Thus the structural distinctions at $\delta_f^*(P)$ and $\delta_f^\dagger(P)$ do not change the exponent in T.

(ii) Linear regret when $C(P) \ne \mathcal{S}$: *If $f \notin \mathcal{F}_{\mathrm{ne}}$ and $\delta < \theta_f(1)$, there exist two weakly communicating kernels P_0, P_1 with $C(P_i) \ne \mathcal{S}$, a common reward r, and a common initial state s_0 such that, for some $c_{f,\delta} > 0$ and all sufficiently large T,*

$$c_{f,\delta} T \le \inf_{\mathfrak{A} \in \mathfrak{A}_{\mathrm{pol}}} \max_{i \in \{0,1\}} \mathbb{E}_{P_i,r,s_0}^{\mathfrak{A}} \operatorname{Reg}_{\mathrm{pol}}(T) \le T.$$

Thus nominal weak communication does not guarantee uniform sublinear regret.

(iii) Saturation: *If $\delta \ge \theta_f(1)$, a policy choosing from $\arg\max_a r(s,a)$ at every state has zero regret for every P. Hence the minimax regret over any nonempty class of nominal kernels is zero. Finite saturation occurs only when both deletion and creation are allowed.*

(iv) Failure of the scalar comparator: *If $f \in \mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$ and $\delta \ge \delta_f^{\mathrm{bell}}(P)$, then some bounded reward has state-dependent robust optimal gain. For such a reward, no common scalar comparator exists, so the regret in* (5) *is not defined.*

Figure 1 summarizes the regret rates in Theorem 5. For the constructions establishing the lower bounds, see the proof of Proposition 16 in Appendix D. The exponents in T are sharp, but the dependence on S, A, and B remains unmatched, even in the common-core regime.

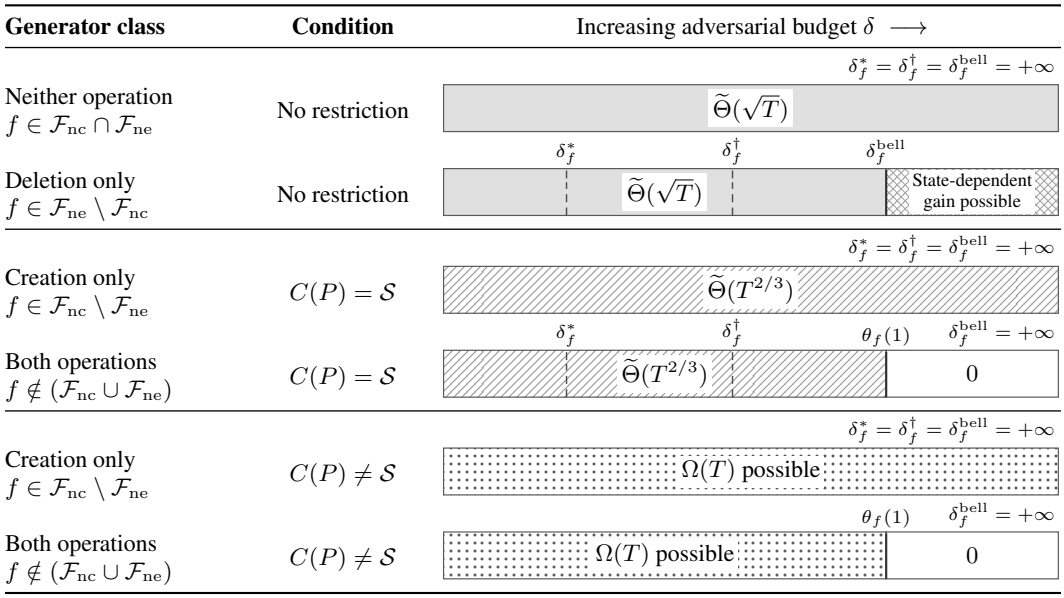

Figure 1: Minimax regret rates from Theorem 5, under $\mathcal{G}(P) \in \mathcal{M}$. Threshold spacing is schematic; finite thresholds may coincide. Dashed marks indicate structural thresholds that leave the matched regret exponent unchanged. For $f \in \mathcal{F}_{\mathrm{ne}}$, the condition $\delta < \delta_f^{\mathrm{bell}}(P)$ already implies $\mathcal{G}(P) \in \mathcal{M}$.

The learner observes transitions under P, but regret evaluates each policy under the worst feasible transition kernel. For $f \in \mathcal{F}_{\mathrm{ne}}$ and $\delta < \delta_f^{\mathrm{bell}}(P)$, Algorithm 3 constructs, with high probability after its initial exploration, policies with a single recurrent class under P. For each such policy, this class contains every state recurrent under any ambiguity kernel. The learner can therefore collect repeated observations at the states needed to bound its robust regret. Estimation errors decrease with the number of visits, and summing their contributions gives the $\widetilde{O}(\sqrt{T})$ bound.

For $f \notin \mathcal{F}_{\mathrm{ne}}$, a policy may never visit a state under P, even though transitions from that state affect its robust value. In the lower-bound instances, obtaining information about such states requires choosing policies with lower robust value. The explore–then–commit bound balances an exploration cost of order m against an estimation loss of order $T/\sqrt{m}$, up to logarithmic factors. Choosing $m \asymp T^{2/3}$ balances these terms. The matching lower bounds show that no learner can achieve a smaller exponent uniformly over these families.

Proposition 1. *Suppose $f \in \mathcal{F}_{\mathrm{ne}}$, $S \geq 3$, and $A \geq 2$. For $B \geq SAC_{f,\delta}$ and all sufficiently large T,*

$$c_{f,\delta}\sqrt{\lfloor S/3 \rfloor AT} \leq \mathfrak{R}_f^{\mathrm{core}}(T;B) \leq \widetilde{O}_{f,\delta}\Big(B^3\sqrt{SAT[S+\log(SAT)]} + B^4S^2A\Big),$$

where $c_{f,\delta}, C_{f,\delta} > 0$ depend only on f and δ.

7 Conclusion and limitations

We identified three thresholds for weak communication of the robust core, weak communication of every ambiguity kernel, and Bellman solvability for every bounded reward. Across the nonempty structural regimes, the minimax regret exponent is $1/2$ for $f \in \mathcal{F}_{\mathrm{ne}}$. For $f \notin \mathcal{F}_{\mathrm{ne}}$, it is $2/3$ when $C(P) = \mathcal{S}$ and $\delta < \theta_f(1)$. Thus, crossing $\delta_f^*(P)$ or $\delta_f^\dagger(P)$ does not change the exponent in T.

The dependence on S, A, and B remains unmatched. In particular, uniform exploration introduces a dependence on $D_0(P)$ that our lower bounds do not establish as necessary. Whether directed exploration can improve this dependence remains open. Any guarantee uniform over initial states must still account for the time needed to reach informative states, including time spent in transient states. We also leave open polynomial-time implementations of BellReady and optimistic robust planning, and explicit bounds on H_{loc} in terms of transition probabilities and structural margins.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims and to assist in writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, or interpret results. Generating synthetic data, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for sourcing/searching for information, editing to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. Critical tools recommended by generative AI were checked against the literature and verified to hold. For example, when $\delta \geq \delta_f^\dagger(P)$, some ambiguity kernels are not weakly communicating, so standard diameter-based span bounds need not apply. Following a suggestion from generative AI, we adapted a rescaling argument from Akian et al. (2015a;b) that reduces the analysis to fixed points of the recession operator. Suggestions made by generative AI for clean notation and proof readability have been verified to be consistent. All relevant literature and sourced information provided by generative AI has been manually surveyed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marianne Akian, Stéphane Gaubert, and Antoine Hochart. Ergodicity conditions for zero-sum games. *Discrete and Continuous Dynamical Systems*, 35(9):3901–3931, 2015a. ISSN 1078-0947. doi: 10.3934/dcds.2015.35.3901. URL https://www.aimsciences.org/article/id/409f25d6-1977-4260-8d5d-a5eab676f5ee.

Marianne Akian, Stéphane Gaubert, and Antoine Hochart. Hypergraph conditions for the solvability of the ergodic equation for zero-sum games. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 5845–5850, 2015b. doi: 10.1109/CDC.2015.7403138.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 133–154. PMLR, 2011. URL https://proceedings.mlr.press/v19/bartok11a.html.

Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26728–26769. Curran Associates, Inc., 2024. doi: 10.52202/079017-0840. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f0bb736ccc8551ef5bcc9165c2a4d9e-Abstract-Conference.html.

Zijun Chen, Shengbo Wang, and Nian Si. Sample complexity of distributionally robust average-reward reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 85402–85463. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7b75da9b61eda40fa35453ee5d077df6-Abstract-Conference.html.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021. doi: 10.1007/s10994-021-05961-4.

Debamita Ghosh, George K. Atia, and Yue Wang. ORVIT: Near-optimal online distributionally robust reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (25):21278–21286, 2026. doi: 10.1609/aaai.v40i25.39273. URL https://ojs.aaai.org/index.php/AAAI/article/view/39273.

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. doi: 10.1287/opre.2023.0694.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22595–22646. PMLR, 2025. URL https://proceedings.mlr.press/v267/he25n.html.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL https://jmlr.org/papers/v11/jaksch10a.html.

Zhipeng Liang, Xiaoteng Ma, Jose Blanchet, Jun Yang, Jiheng Zhang, and Zhengyuan Zhou. Single-trajectory distributionally robust reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29644–29666. PMLR, 2024. URL https://proceedings.mlr.press/v235/liang24d.html.

Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust Markov decision processes. *Mathematics of Operations Research*, 41(4):1325–1353, 2016. doi: 10.1287/moor.2016.0779. URL https://doi.org/10.1287/moor.2016.0779.

Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12528–12580. Curran Associates, Inc., 2024. doi: 10.52202/079017-0399. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/170dc3e41f2d03e327e04dbab0fccbfb-Abstract-Conference.html.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

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. PMLR, 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32211–32224. Curran Associates, Inc., 2022. doi: 10.52202/068431-2334. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d01bda31bbcd780774ff15b534e03c40-Abstract-Conference.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994. ISBN 978-0-471-61977-2. doi: 10.1002/9780470316887.

Zachary Roch, George Atia, and Yue Wang. Model-free robust average-reward reinforcement learning with sample complexity analysis, 2026. URL https://arxiv.org/abs/2505.12462. Accepted at the 43rd International Conference on Machine Learning (ICML 2026).

Zachary Andrew Roch, George K. Atia, and Yue Wang. A reduction framework for distributionally robust reinforcement learning under average reward. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51809–51850. PMLR, 2025. URL https://proceedings.mlr.press/v267/roch25a.html.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024. URL https://jmlr.org/papers/v25/22-1482.html.

Shuo Sun, Meng Qi, and Zuo-Jun Max Shen. Online MDP with prototypes information: A robust adaptive approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):20717–20724, 2025. doi: 10.1609/aaai.v39i19.34283. URL https://ojs.aaai.org/index.php/AAAI/article/view/34283.

Ambuj Tewari and Peter L. Bartlett. Bounded parameter Markov decision processes with average reward criterion. In Nader H. Bshouty and Claudio Gentile (eds.), *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pp. 263–277, Berlin, Heidelberg, 2007. Springer. doi: 10.1007/978-3-540-72927-3_20.

Shengbo Wang and Nian Si. Bellman optimality of average-reward robust Markov decision processes with a constant gain, 2025. URL https://arxiv.org/abs/2509.14203.

Shengbo Wang and Nian Si. Non-rectangular average-reward robust MDPs: Optimal policies and their transient values, 2026.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward Markov decision processes. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7fde8314b66c570c210f30cbd6c1735d-Abstract-Conference.html.

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. URL https://www.jmlr.org/papers/v25/23-0526.html.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7193–7206. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/3a4496776767aaa99f9804d0905fe584-Abstract.html.

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. doi: 10.1613/jair.1.15451. URL https://www.jair.org/index.php/jair/article/view/15451.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. doi: 10.1287/moor.1120.0566.

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. Curran Associates, Inc., 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/df15617fd6d6f78fcd485401d0598761-Abstract-Conference.html.

Yang Xu, Swetha Ganesh, and Vaneet Aggarwal. Efficient Q-Learning and actor–critic methods for robust average-reward reinforcement learning. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7613–7650. PMLR, 17–21 Aug 2026.

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.

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. PMLR, 2023. URL https://proceedings.mlr.press/v195/zhang23b.html.

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. Curran Associates, Inc., 2024. doi: 10.52202/079017-1055. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3acbe9dc3a1e8d48a57b16e9aef91879-Abstract-Conference.html.

$f(t)$	$f(0)$	$f'(\infty)$	$\theta_f(x)$	$q_\star(f,\delta)$	Class
$t\log t$ (KL)	0	$+\infty$	$-\log(1-x)$	$1-e^{-\delta}$	$\mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$
$-\log t + t - 1$ (reverse KL)	$+\infty$	1	$\begin{cases} 0, & x = 0, \\ +\infty, & x > 0 \end{cases}$	0	$\mathcal{F}_{\mathrm{nc}} \setminus \mathcal{F}_{\mathrm{ne}}$
$\frac{1}{2}(t-1)^2$ (Pearson)	$\frac{1}{2}$	$+\infty$	$\frac{x}{2(1-x)}$	$\frac{2\delta}{1+2\delta}$	$\mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$
$\frac{1}{2}\lvert t-1\rvert$ (TV)	$\frac{1}{2}$	$\frac{1}{2}$	x	$\min\{\delta, 1\}$	both operations
$(\sqrt{t}-1)^2$ (squared Hellinger)	1	1	$2(1-\sqrt{1-x})$	$\begin{cases} \delta - \delta^2/4, & \delta < 2, \\ 1, & \delta \geq 2 \end{cases}$	both operations

Table 1: Support geometry for the displayed generators in $D_f(p\|q)$, with nominal row q and $0 < \delta < \infty$. The Pearson row uses one-half the usual χ^2 divergence; squared Hellinger uses $\sum_i(\sqrt{p_i} - \sqrt{q_i})^2$. Here $q_\star = \sup\{x \in [0,1] : \theta_f(x) \leq \delta\}$. Values at $x = 1$ are interpreted by limits. The classes $\mathcal{F}_{\mathrm{ne}}$ and $\mathcal{F}_{\mathrm{nc}}$ prohibit transition creation and deletion, respectively. "Both operations" means that $f(0)$ and $f'(\infty)$ are finite. Deleting a successor set with nominal probability $0 < x < 1$ is feasible exactly when $\theta_f(x) \leq \delta$.

A Extended related work

Foundational treatments of rectangular robust dynamic programming include Nilim & El Ghaoui (2005) and Iyengar (2005). Discounted robust RL has generative and offline guarantees (Wang et al., 2024b; Shi & Chi, 2024; Panaganti & Kalathil, 2022). For single-trajectory learning, Wang & Zou (2021) give finite-time bounds for contamination uncertainty under ergodicity and coverage conditions, while Liang et al. (2024) establish asymptotic convergence for f-divergence uncertainty. In finite-horizon online learning, Lu et al. (2024) use a vanishing-value condition and He et al. (2025) impose visitation conditions; ORVIT removes these specific restrictions (Ghosh et al., 2026). Sun et al. (2025) instead study known transition prototypes and shrinking ambiguity sets, with regret measured under the true kernel.

Average reward raises an additional difficulty: changes to the transition graph can make the optimal worst-case average reward depend on the starting state. For f-divergence balls, standard divergence properties determine which graph changes are possible. A nominally absent transition can receive positive probability at any positive finite radius precisely when $f'(\infty) < \infty$. Deleting successors carrying total nominal probability $0 < x < 1$ has minimum divergence cost $\theta_f(x)$. We use these facts to derive three exact radii characterizing a common weakly communicating core, weak communication of every ambiguity kernel, and scalar robust Bellman solvability for every bounded reward. We then connect these radii to online regret rates. The support-creation mechanism also appears in the TV hardness construction of Lu et al. (2024): robust evaluation can depend on states inaccessible to nominal data collection. Our results characterize how this mechanism affects learning in the continuing average-reward setting.

Our Bellman analysis builds on the nonlinear ergodicity theory of Akian et al. (2015a;b). In the finite-state setting considered here, this theory characterizes scalar Bellman solvability for every bounded reward. The obstruction is a pair of nonempty, disjoint state sets: the controller can prevent exit from one against any adversary, while the adversary can prevent exit from the other against any control policy. For robust MDPs, Wang & Si (2025) establish scalar Bellman solvability under communication conditions. For state–action rectangular f-divergence balls, we use transition deletion and creation costs to determine exactly when the obstruction arises. This yields the radius $\delta_f^{\mathrm{bell}}(P)$ and distinguishes it from the two communication thresholds.

Existing robust average-reward work includes optimality and convergence results (Tewari & Bartlett, 2007; Wang et al., 2024c; Grand-Clément et al., 2026), and finite-sample analyses of policy evaluation and control under generative access (Roch et al., 2025; Chen et al., 2025; Roch et al., 2026; Xu et al., 2025; 2026; Yang et al., 2026). Yang et al. (2026) characterize minimax generative-model complexity for state–action rectangular TV balls, assuming every stationary policy is unichain under every admissible kernel, and identify learning without a simulator as future work. We address this direction for policy-valued regret from one nominal trajectory, with matching horizon exponents on

parameter-bounded families. A generative model allows the learner to sample any state–action pair directly. With a single trajectory, the learner must first reach the state.

Related continuing online results use different observation models and regret criteria. Lim et al. (2016) allow unknown stochastic and adversarial rows, with unknown uncertainty sets within prescribed nested families. Their average-reward extension assumes communicating and unichain structure for every admissible kernel choice and bounds regret based on accumulated realized rewards. Wang & Si (2026) connect realized-reward regret to robust optimality for a prescribed, possibly nonrectangular ambiguity set, when the adversary commits to a stationary kernel. In our setting, the learner observes one trajectory from an unknown nominal kernel. At each step, regret measures the difference between the optimal robust gain and that of the announced stationary policy, both evaluated at the current state.

Nominal benchmarks include UCRL2 (Jaksch et al., 2010), model-free algorithms with optimal $\sqrt{T}$ horizon dependence (Zhang & Xie, 2023), and tractable algorithms with a minimax-optimal leading regret term in terms of the optimal bias span (Boone & Zhang, 2024). Generative sample-complexity guarantees cover uniformly ergodic MDPs (Wang et al., 2024a) and weakly communicating and general multichain MDPs (Zurek & Chen, 2024). Our $T^{2/3}$ lower bound uses an explicit MDP pair in which distinguishing the optimal policies requires costly exploration, analogous to the information-acquisition tradeoff in stochastic partial monitoring (Bartók et al., 2011).

B Notation not carried in the main text

B.1 Confidence sets and the topology certificate

Fix $\alpha \in (0,1)$, let $\bar{\alpha} := \alpha/3$ and $\bar{n} := \max\{1, n\}$, and define

$$\beta_{\bar{\alpha}}(n) := \min\Big\{2, \sqrt{2\bar{n}^{-1}\log(c_{\bar{\alpha}}\bar{n}(\bar{n}+1))}\Big\}, \qquad c_{\bar{\alpha}} := \frac{SA(2^S-2)}{\bar{\alpha}}.$$

The case $S = 1$ is trivial, every row being the point mass: we set $\beta_{\bar{\alpha}} \equiv 0$ there, and the formula for $c_{\bar{\alpha}}$ is read for $S \geq 2$. Let $N_k(s,a)$ be the visits to (s,a) before episode k and let $\widehat{P}_{k,s,a}$ be its empirical row, uniform when $N_k(s,a) = 0$. The per-episode balls, their intersections, and the simultaneous confidence event are

$$\begin{aligned}
\mathcal{B}_j(s,a) &:= \big\{q \in \Delta(\mathcal{S}) : \|q - \widehat{P}_{j,s,a}\|_1 \leq \beta_{\bar{\alpha}}(N_j(s,a))\big\},\\
\mathcal{C}_k(s,a) &:= \bigcap_{j \leq k} \mathcal{B}_j(s,a),\\
\mathcal{E}_{\mathrm{conf}} &:= \{P_{s,a} \in \mathcal{C}_k(s,a)\ \forall k,\ \forall (s,a)\},
\end{aligned}$$

together with the bracketing graphs

$$\begin{aligned}
H_k^-(s,a) &:= \{s' : \inf_{q \in \mathcal{C}_k(s,a)} \theta_f(q(s')) > \delta\},\\
H_k^+(s,a) &:= \{s' : \sup_{q \in \mathcal{C}_k(s,a)} \theta_f(q(s')) > \delta\}.
\end{aligned}$$

For $B \neq \varnothing$, let $G|_B(s,a) := G(s,a) \cap B$ for $s \in B$ and let $\mathcal{M}(B)$ be the weakly communicating graphs on B. Define

$$H_k^+(s,a) \subseteq B \quad (s \in B, a \in \mathcal{A}), \qquad H_k^-|_B \in \mathcal{M}(B), \tag{8}$$

and call B a certified recurrent region when both conditions hold. Define

$$\begin{aligned}
\mathfrak{B}_k &:= \{B \neq \varnothing : \text{condition (8) holds}\},\\
\tau_{\mathrm{topo}} &:= \inf\{t_k : \mathfrak{B}_k \neq \varnothing\},
\end{aligned}$$

fixing some $B_k \in \mathfrak{B}_k$ for $t_k \geq \tau_{\mathrm{topo}}$. Since $\mathcal{C}_k$ is nonincreasing in k, the graph H_k^- is nondecreasing and H_k^+ nonincreasing, so a valid certificate stays valid.

Let π_0 be uniform and define the global navigation time

$$D_0(P) := 1 + \max_{\substack{x \in \mathcal{S},\ s \in C(P)\\ a \in \mathcal{A}}} \mathbb{E}_P^{\pi_0}[\inf\{t : (S_t, A_t) = (s,a)\} \mid S_0 = x].$$

Thus D_0 controls both entrance into $C(P)$ and subsequent coverage of its state–action rows. In particular, with $\tau_C := \inf\{t : S_t \in C(P)\}$,

$$\max_{x\in\mathcal{S}} \mathbb{E}_P^{\pi_0}[\tau_C \mid S_0 = x] \le D_0(P) - 1.$$

For an edge e, let $I_e(\epsilon) := [0,1] \cap [P(e) - \epsilon, P(e) + \epsilon]$ and define

$$\begin{aligned}
\epsilon_0(P) &:= \begin{cases} +\infty, & C(P) = \mathcal{S}, \\ \sup\{\epsilon \in (0,1] : \sup_{x\in[0,\epsilon]} \theta_f(x) \le \delta\}, & C(P) \neq \mathcal{S}, \end{cases} \\
\epsilon_e(P) &:= \sup\{\epsilon > 0 : \inf_{x\in I_e(\epsilon)} \theta_f(x) > \delta\}, \\
\epsilon_{s,a}(P) &:= \epsilon_0(P) \wedge \min_{s'\in\mathcal{G}_\wedge(P)(s,a)} \epsilon_{(s,a,s')}(P), \\
\epsilon_{\mathrm{top}}(P) &:= \min_{s\in C(P), a\in\mathcal{A}} \epsilon_{s,a}(P), \\
\Delta_\theta(P) &:= \min_{\substack{s\in C(P), a\in\mathcal{A} \\ s'\in\mathcal{G}_\wedge(P)(s,a)}} \{\theta_f(P(s' \mid s,a)) - \delta\}.
\end{aligned}$$

We use $\sup\varnothing := 0$, $\min\varnothing := +\infty$, and $1/(+\infty)^2 := 0$. Monotonicity of θ_f gives

$$\epsilon_0(P) = q_\star \quad (C(P) \neq \mathcal{S}), \qquad\qquad \epsilon_e(P) = P(e) - q_\star \tag{9}$$

for the edges appearing in the core regime.

The per-row sample counts and the local modulus of θ_f referred to in the main text are

$$\begin{aligned}
n_{s,a}^{\mathrm{top}}(P,\alpha) &:= \min\{n \ge 1 : 2\beta_{\bar\alpha}(n) < \epsilon_{s,a}(P)\}, \\
M_{\mathrm{top}}(P,\alpha) &:= \textstyle\sum_{s\in C(P),\, a\in\mathcal{A}} n_{s,a}^{\mathrm{top}}(P,\alpha), \\
\eta_f &:= \tfrac{1}{2}\big(1 - q_\star(f,\delta)\big) \wedge \tfrac{1}{2}, \\
L_f &:= 1 \vee \sup_{x\neq y \text{ in } [q_\star, q_\star+\eta_f]\cap[0,1)} \left|\frac{\theta_f(x) - \theta_f(y)}{x - y}\right|,
\end{aligned}$$

with $\min\varnothing := +\infty$. Both depend only on (f,δ) and not on P; Lemma 4 implies $L_f < \infty$ whenever $f(0) < \infty$, and $L_f := 1$ is used when $f(0) = +\infty$, where $\Delta_\theta \equiv +\infty$ makes the constant immaterial.

B.2 Recurrent restriction and Bellman-stability quantities

Define

$$R(P) := \begin{cases} C(P), & f \in \mathcal{F}_{\mathrm{ne}}, \\ \mathcal{S}, & f \notin \mathcal{F}_{\mathrm{ne}}, \end{cases} \qquad d_R := |R(P)|,$$

and let P_R denote the restriction of P to rows and successors in $R(P)$. For $n \in \mathbb{N}_0$, define the uniform-exploration clock

$$\begin{aligned}
\mathfrak{t}_P(n,\alpha) &:= \begin{cases} 0, & n = 0, \\ 2D_0(P) d_R A n \left\lceil \log_2 \dfrac{6 d_R A n}{\alpha} \right\rceil, & n \ge 1, \end{cases} \\
n_T(P,\alpha) &:= \max\{n \in \mathbb{N}_0 : \mathfrak{t}_P(n,\alpha) \le \lceil T^{2/3} \rceil\}.
\end{aligned}$$

With $\|Q - P_R\|_{\max,1} := \max_{s\in R(P),a} \|Q_{s,a} - P_{R,s,a}\|_1$, let $\epsilon_{\mathrm{wc}}(P_R)$ be the largest radius on which $\delta < \delta_f^{\mathrm{bell}}(Q)$ persists, restricted to the nominal support face when $f \in \mathcal{F}_{\mathrm{ne}}$, where the nominal support face is defined by the closed zero constraints $Q(s' \mid s,a) = 0$ whenever $P_R(s' \mid s,a) = 0$. Let

$$\begin{aligned}
p_{\min,R}(P) &:= \min\{P(s' \mid s,a) > 0 : s, s' \in R(P), a \in \mathcal{A}\}, \\
\bar\epsilon_{\mathrm{wc}}(P_R) &:= \tfrac{1}{2}\min\{1, \epsilon_{\mathrm{wc}}(P_R), p_{\min,R}(P)\}, \\
\bar\epsilon_{\mathrm{face}}(P_R) &:= \tfrac{1}{4}\min\{\bar\epsilon_{\mathrm{wc}}(P_R), p_{\min,R}(P)\}.
\end{aligned}$$

Capping the radius by $p_{\min,R}(P)$ keeps every nominally positive coordinate of a kernel within distance $\bar\epsilon_{\mathrm{wc}}(P_R)$ of P_R at least $3p_{\min,R}(P)/4$: an individual coordinate changes by at most half the

row ℓ_1 distance, and $\bar{\epsilon}_{\mathrm{wc}}(P_R) \le \frac{1}{2} p_{\min,R}(P)$. Let $\mathcal{K}(P_R)$ be the corresponding closed ball, imposing the zero constraints above when $f \in \mathcal{F}_{\mathrm{ne}}$, and let $\mathcal{K}_{\mathrm{face}}(P_R)$ impose the same zero constraints in all cases, with radius $\bar{\epsilon}_{\mathrm{face}}(P_R)$.

For kernels on $R(P)$ define

$$\mathfrak{W}_{\mathrm{bell}}(Q) := \{(I,J) : I, J \neq \varnothing,\ I \cap J = \varnothing,\ J \text{ closable in } \mathcal{G}(Q)\},$$

$$m_{\mathrm{bell}}(Q) := \begin{cases} +\infty, & f \notin \mathcal{F}_{\mathrm{ne}} \text{ or } \mathfrak{W}_{\mathrm{bell}}(Q) = \varnothing, \\ \min\limits_{(I,J)\in\mathfrak{W}_{\mathrm{bell}}(Q)} \max\limits_{s\in I, a\in\mathcal{A}} [Q(I^c \mid s,a) - q_\star], & f \in \mathcal{F}_{\mathrm{ne}}, \end{cases}$$

$$\Delta_{\mathrm{bell}}(P_R) := \min\{1, p_{\min,R}(P), m_{\mathrm{bell}}(P_R)\}.$$

Over $\mathfrak{U}_1 := \{u \in \mathbb{R}^{R(P)} : \min_x u(x) = 0, \mathrm{sp}(u) \le 1\}$, define

$$\varpi^{P_R}_{f,\delta}(\epsilon) := \sup_{\substack{Q,Q'\in\mathcal{K}(P_R), u\in\mathfrak{U}_1 \\ s\in R(P), a\in\mathcal{A}, \|Q_{s,a} - Q'_{s,a}\|_1 \le \epsilon}} |\sigma^f_\delta(Q_{s,a}, u) - \sigma^f_\delta(Q'_{s,a}, u)|,$$

$$\varpi^{P_R}_{f,\delta,\mathrm{face}}(\epsilon) := \sup_{\substack{Q,Q'\in\mathcal{K}_{\mathrm{face}}(P_R), u\in\mathfrak{U}_1 \\ s\in R(P), a\in\mathcal{A}, \|Q_{s,a} - Q'_{s,a}\|_1 \le \epsilon}} |\sigma^f_\delta(Q_{s,a}, u) - \sigma^f_\delta(Q'_{s,a}, u)|,$$

$$K_{\mathrm{face}}(P_R) := \sup_{\epsilon>0} \varpi^{P_R}_{f,\delta,\mathrm{face}}(\epsilon)/\epsilon,$$

$$H_{\mathrm{loc}}(P_R) := \sup_{Q\in\mathcal{K}(P_R), \gamma\in(0,1)} \mathrm{sp}(v_{Q,\gamma}),$$

where $v_{Q,\gamma}$ is the normalized fixed point of $\mathcal{L}^{\mathrm{rob}}_{Q,\gamma}$ on $R(P)$.

B.3 Optimistic operators and occupation weights

For $t_k \ge \tau_{\mathrm{topo}}$ the optimistic robust operator and its policy-specific counterpart are

$$\mathcal{L}_k u(s) := \max_a \big\{ r(s,a) + \sup_{q\in\mathcal{C}_k(s,a)} \sigma^f_\delta(q,u) \big\},$$

$$\mathcal{L}_{k,\pi} u(s) := r(s,\pi(s)) + \sup_{q\in\mathcal{C}_k(s,\pi(s))} \sigma^f_\delta(q,u).$$

Let $(g_k, h_k) \in \mathrm{Eig}(\mathcal{L}_k)$, let π_k attain the maximum in $\mathcal{L}_k h_k$, and let $\bar{\pi}_k$ be a completion of π_k supplied by Lemma 14. The two exact local losses, the adversarial stationary weights, and the induced occupation measure are

$$\zeta_k(s) := [\mathcal{L}_k h_k](s) - [\mathcal{L}_{k,\bar{\pi}_k} h_k](s),$$

$$b_k(s,a) := \sup_{q\in\mathcal{C}_k(s,a)} \sigma^f_\delta(q,h_k) - \sigma^f_\delta(P_{s,a}, h_k),$$

$$\widetilde{Q}^*_k \in \operatorname*{arg\,min}_{\widetilde{Q}\in\mathcal{U}^P(\delta)} g^{\bar{\pi}_k}_{\widetilde{Q}},$$

$$\mu_k \in \Delta(\mathcal{S}) \text{ with } \mu_k^\top (\widetilde{Q}^*_k)_{\bar{\pi}_k} = \mu_k^\top,$$

$$\rho_k(s,a) := \mu_k(s) \mathbf{1}\{a = \bar{\pi}_k(s)\}.$$

On $\mathcal{E}_{\mathrm{conf}}$ we have $b_k \ge 0$, and ζ_k vanishes wherever $\bar{\pi}_k$ retains the greedy action.

B.4 Two-point quantities and coverage

For nominal kernels P_0, P_1 whose robust policy gains are scalar, meaning $g^\pi_{\mathrm{rob},P_i} \in \mathbb{R}\mathbf{1}$ for every $\pi \in \Pi$, let

$$\ell_i(\pi) := g^{\mathrm{rob}}(P_i) - g^\pi_{\mathrm{rob},P_i},$$

$$\Delta_{\mathrm{dec}}(P_0, P_1) := \inf_{\pi\in\Pi} \{\ell_0(\pi) + \ell_1(\pi)\},$$

$$\mathcal{I}_T := \mathrm{KL}\big(\mathbb{P}^{\mathfrak{A},T}_{P_0} \,\big\|\, \mathbb{P}^{\mathfrak{A},T}_{P_1}\big).$$

Here $\mathbb{P}_{P_i}^{\mathfrak{A},T}$ is the law of (U, H_T) under P_i, including the learner's private random seed. When only the robust optimal gains are scalar, define instead

$$\ell_i(\pi, x) := g^{\mathrm{rob}}(P_i) - g^{\pi}_{\mathrm{rob},P_i}(x),$$
$$\Delta^{\mathrm{st}}_{\mathrm{dec}}(P_0, P_1) := \inf_{\pi\in\Pi,\, x\in\mathcal{S}} \{\ell_0(\pi, x) + \ell_1(\pi, x)\}.$$

The superscript records that the separation is uniform in the current state. If every policy gain is scalar, then $\Delta^{\mathrm{st}}_{\mathrm{dec}} = \Delta_{\mathrm{dec}}$.

For $q \in [0, 1]$ the smallest robustly feasible mass on the first coordinate is

$$\underline{p}_f(q) := \min\{p \in [0, 1] : D_f((p, 1-p)\|(q, 1-q)) \le \delta\},$$

so that $\underline{p}_f(q) = 0$ exactly when $\theta_f(q) \le \delta$. Finally, whenever $\delta < \delta^*_f(P)$, the coverage coefficient is

$$c_{\mathrm{cov}}(P) := \max_{\substack{\pi\in\Pi_{\mathrm{d}}:\\ \pi \text{ unichain in } \mathcal{G}_\wedge(P)}} \max_{\widetilde{Q}\in\mathcal{U}^P} \max_{s\in\mathcal{S}} \frac{\mu^{\pi}_{\widetilde{Q}}(s)}{\mu^{\pi}_{P}(s)},$$

with $0/0 := 0$ and $x/0 := +\infty$ for $x > 0$. No result in this paper assumes $c_{\mathrm{cov}}(P) < \infty$; it appears only to record that Proposition 14 produces instances where it is infinite.

C Robust support geometry and Bellman structure

For $q \in \Delta(\mathcal{S})$ and $B \subseteq \mathcal{S}$, write $q(B) := \sum_{i\in B} q_i$. Recall

$$\theta_f(x) := xf(0) + (1-x)f\left(\frac{1}{1-x}\right), \qquad x \in [0, 1),$$

with $\theta_f(1) = f(0) + f'(\infty)$.

Lemma 2. *For every $q \in \Delta(\mathcal{S})$ and every proper $B \subsetneq \mathcal{S}$,*

$$\inf_{\substack{p\in\Delta(\mathcal{S})\\ p(B)=0}} D_f(p\|q) = \theta_f(q(B)).$$

If $q(B) < 1$, the infimum is attained by

$$p^B_i = \begin{cases} 0, & i \in B, \\ q_i/(1-q(B)), & i \notin B. \end{cases}$$

Proof. Assume first that $x := q(B) < 1$, and fix an arbitrary $p \in \Delta(\mathcal{S})$ satisfying $p(B) = 0$. Partition B^c into

$$Z := \{i \in B^c : q_i > 0\}, \qquad N := \{i \in B^c : q_i = 0\}.$$

For $i \in B$ we have $p_i = 0$. Hence the total contribution of the coordinates in B is

$$\sum_{i\in B: q_i>0} q_i f(0) + \sum_{i\in B: q_i=0} 0 = q(B)f(0) = xf(0).$$

Define

$$m := \sum_{i\in Z} p_i.$$

Because $p(B) = 0$ and p has total mass one,

$$\sum_{i\in N} p_i = 1 - m.$$

Every coordinate in N has zero nominal mass, so the singular-mass convention gives the contribution

$$\sum_{i\in N} p_i f'(\infty) = (1-m)f'(\infty).$$

It remains to lower-bound the contribution on Z. Since $\sum_{i\in Z} q_i = q(B^c) = 1 - x > 0$, the numbers

$$\alpha_i := \frac{q_i}{1-x}, \qquad i \in Z,$$

are nonnegative and sum to one. Convexity of f therefore gives

$$\begin{aligned}\sum_{i\in Z} q_i f\left(\frac{p_i}{q_i}\right) &= (1-x)\sum_{i\in Z}\alpha_i f\left(\frac{p_i}{q_i}\right)\\ &\geq (1-x) f\left(\sum_{i\in Z}\alpha_i \frac{p_i}{q_i}\right)\\ &= (1-x) f\left(\frac{m}{1-x}\right).\end{aligned}$$

Set

$$t := \frac{m}{1-x}, \qquad r := \frac{1}{1-x}.$$

Because $0 \leq m \leq 1$, we have $0 \leq t \leq r$. If $t < r$, convexity implies that every secant slope is bounded above by the recession slope:

$$\frac{f(r) - f(t)}{r - t} \leq f'(\infty). \tag{10}$$

Rearranging inequality (10) gives

$$f(t) + (r-t) f'(\infty) \geq f(r).$$

The resulting inequality also holds with equality when $t = r$. Since

$$(1-x)(r-t) = 1 - m,$$

multiplication by $1 - x$ yields

$$\begin{aligned}&(1-x) f\left(\frac{m}{1-x}\right) + (1-m) f'(\infty)\\ &\geq (1-x) f\left(\frac{1}{1-x}\right).\end{aligned}$$

Combining the contributions on B, Z, and N proves that every feasible p satisfies

$$D_f(p\|q) \geq x f(0) + (1-x) f\left(\frac{1}{1-x}\right) = \theta_f(x).$$

Now consider the proposed distribution p^B. It has no singular mass, and for every $i \in B^c$ with $q_i > 0$ its likelihood ratio is

$$\frac{p_i^B}{q_i} = \frac{1}{1-x}.$$

Consequently,

$$\begin{aligned}D_f(p^B\|q) &= \sum_{i\in B: q_i>0} q_i f(0) + \sum_{i\in B^c: q_i>0} q_i f\left(\frac{1}{1-x}\right)\\ &= x f(0) + (1-x) f\left(\frac{1}{1-x}\right)\\ &= \theta_f(x).\end{aligned}$$

Thus the lower bound is attained and is the required infimum.

It remains to treat $x = 1$. In this case $q(B) = 1$, so $q_i = 0$ for every $i \in B^c$. Since $p(B) = 0$, every feasible p places all of its mass on B^c. The coordinates in B contribute

$$\sum_{i\in B:\, q_i>0} q_i f(0) = q(B) f(0) = f(0),$$

and the unit mass placed on zero-q coordinates contributes $f'(\infty)$. Hence every feasible p has divergence

$$D_f(p\|q) = f(0) + f'(\infty) = \theta_f(1),$$

which proves the claim also when $q(B) = 1$. □

Lemma 3. *Fix i with $q_i = 0$. If $f'(\infty) = +\infty$, then every p with $p_i > 0$ satisfies $D_f(p\|q) = +\infty$. If $f'(\infty) < +\infty$, then, for every $\delta > 0$, there exists $p \in \mathcal{U}^q(\delta)$ with $p_i > 0$.*

Proof. Assume first that $f'(\infty) = +\infty$ and let $p_i > 0$. Since $q_i = 0$, the ith coordinate is singular with respect to q and contributes

$$p_i f'(\infty) = +\infty$$

to $D_f(p\|q)$. Thus $D_f(p\|q) = +\infty$, so such a p cannot belong to any finite-radius ambiguity set.

Now assume that $f'(\infty) < +\infty$. For $\epsilon \in (0, 1)$ define

$$p^\epsilon := (1 - \epsilon)q + \epsilon e_i.$$

Then $p^\epsilon \in \Delta(\mathcal{S})$ and $p^\epsilon_i = \epsilon > 0$. For every coordinate j with $q_j > 0$ we have

$$\frac{p^\epsilon_j}{q_j} = 1 - \epsilon,$$

whereas the mass ϵ at coordinate i is singular. Therefore

$$D_f(p^\epsilon\|q) = \sum_{j:q_j>0} q_j f(1 - \epsilon) + \epsilon f'(\infty) = f(1 - \epsilon) + \epsilon f'(\infty).$$

Because a finite convex function is continuous on the interior of its domain and f is finite on $(0, \infty)$,

$$\lim_{\epsilon\downarrow 0} f(1 - \epsilon) = f(1) = 0.$$

The finiteness of $f'(\infty)$ also gives $\epsilon f'(\infty) \to 0$. Hence

$$\lim_{\epsilon\downarrow 0} D_f(p^\epsilon\|q) = 0.$$

For the prescribed $\delta > 0$, choose $\epsilon > 0$ sufficiently small that this divergence is at most δ. Then $p^\epsilon \in \mathcal{U}^q(\delta)$ and $p^\epsilon_i > 0$, proving the second claim. □

C.1 Deletion-cost monotonicity and continuity

Lemma 4. *For every $f \in \mathcal{F}$ the map θ_f is nondecreasing on $[0, 1]$, with $\theta_f(0) = 0$. If moreover $f(0) < +\infty$, then θ_f is real valued, convex and continuous on $[0, 1)$.*

Proof. The value at 0 is $\theta_f(0) = 0 \cdot f(0) + f(1) = 0$ under the convention $0 \cdot (+\infty) := 0$.

Fix $0 < x < y < 1$ and set $p_y := (0, 1)$, $q_y := (y, 1 - y)$. Let K be the Markov kernel on two points that sends the first coordinate to the first coordinate with probability x/y and to the second coordinate otherwise, and sends the second coordinate to the second coordinate with probability one. Then

$$p_y K = (0, 1), \qquad q_y K = (x, 1 - x).$$

For each output coordinate j define

$$w_{ij} := \frac{q_{y,i} K(j \mid i)}{(q_y K)_j}, \qquad i \in \{1, 2\}.$$

The two weights are nonnegative and sum to one, and

$$\frac{(p_y K)_j}{(q_y K)_j} = \sum_{i=1}^{2} w_{ij} \frac{p_{y,i}}{q_{y,i}}.$$

Convexity of f, followed by summation over j, therefore gives

$$\begin{aligned} D_f(p_y K\|q_y K) &\le \sum_{j=1}^{2} (q_y K)_j \sum_{i=1}^{2} w_{ij} f\left(\frac{p_{y,i}}{q_{y,i}}\right) \\ &= \sum_{i=1}^{2} q_{y,i} f\left(\frac{p_{y,i}}{q_{y,i}}\right) = D_f(p_y\|q_y). \end{aligned} \tag{11}$$

This is the data-processing inequality for the present two-point f-divergence experiment. Lemma 2 applied with $B = \{1\}$ identifies $D_f((0,1)\|(z,1-z)) = \theta_f(z)$ for every $z \in [0,1]$. Hence inequality (11) gives

$$\theta_f(x) = D_f(p_y K \| q_y K) \le D_f(p_y \| q_y) = \theta_f(y).$$

For $x = 0$, convexity gives $\theta_f(y) \ge f(y \cdot 0 + (1-y)/(1-y)) = f(1) = 0 = \theta_f(0)$. The case $y = 1$ follows by taking $y' \uparrow 1$ and using the definition $\theta_f(1) := \lim_{y' \uparrow 1} \theta_f(y')$.

Now suppose $f(0) < +\infty$. Then $x \mapsto xf(0)$ is affine and finite. For $t \in (0,1]$ the map $t \mapsto tf(1/t)$ is the perspective of the convex function f evaluated at $(1,t)$, hence convex and finite; composing with the affine map $x \mapsto 1-x$ shows that $x \mapsto (1-x)f(1/(1-x))$ is convex and finite on $[0,1)$. Thus θ_f is convex and finite there, hence continuous on $(0,1)$ and, by the computation of $\theta_f(0)$ above together with $\lim_{x \downarrow 0}(1-x)f(1/(1-x)) = f(1) = 0$, also continuous at 0. □

Lemma 5. *Let $q_\star = q_\star(f,\delta)$ be as in equation (1). $\theta_f(x) \le \delta$ if and only if $x \le q_\star$; moreover $q_\star = 0$ if and only if $f \in \mathcal{F}_{\mathrm{nc}}$, and $q_\star < 1$ if and only if $\delta < \theta_f(1)$. If $f(0) < +\infty$ and $q_\star < 1$, then $\theta_f(q_\star) = \delta$.*

Proof. The two monotonicity statements are immediate from Lemma 4 and the definition of $q_\star$ as a supremum. If $f(0) = +\infty$ then $\theta_f(x) = +\infty > \delta$ for every $x > 0$, so $q_\star = 0$; conversely if $f(0) < +\infty$ then θ_f is continuous at 0 with $\theta_f(0) = 0 < \delta$, so $\theta_f(x) \le \delta$ for all small $x > 0$ and $q_\star > 0$.

For the third claim, suppose first $q_\star < 1$. Then there is $x \in (q_\star, 1)$ with $\theta_f(x) > \delta$, and monotonicity gives $\theta_f(1) \ge \theta_f(x) > \delta$. Conversely suppose $\delta < \theta_f(1)$. If $f(0) = +\infty$ then $q_\star = 0 < 1$ by the previous paragraph. If $f(0) < +\infty$ then $\theta_f(x) \to \theta_f(1)$ as $x \uparrow 1$, by the definition of $\theta_f(1)$ as that limit, so some $x < 1$ has $\theta_f(x) > \delta$ and hence $q_\star \le x < 1$.

Finally let $f(0) < +\infty$ and $q_\star < 1$. Then θ_f is continuous on $[0,1)$ by Lemma 4. Taking $x_n \uparrow q_\star$ with $\theta_f(x_n) \le \delta$ gives $\theta_f(q_\star) \le \delta$, and taking $x_n \downarrow q_\star$ with $x_n < 1$ and $\theta_f(x_n) > \delta$ gives $\theta_f(q_\star) \ge \delta$.

This also settles the endpoint in the first claim. If $q_\star = 0$, then $\theta_f(0) = 0 \le \delta$; if $q_\star = 1$, monotonicity and $\delta \ge \theta_f(1)$ give $\theta_f(x) \le \delta$ for every x. In the remaining case, continuity gives $\theta_f(q_\star) = \delta$ as shown above. □

C.2 Structure of recurrent regions

Lemma 6. *Let $G \in \mathcal{M}$. Then the nonempty closed sets of G are closed under intersection and have a minimum element $I_{\min}(G)$, and*

$$C(G) = I_{\min}(G).$$

Consequently: (i) $C(G)$ is closed in G; (ii) every nonempty closed set of G contains $C(G)$; (iii) any two states of $C(G)$ are mutually reachable in G; (iv) every state of $T(G)$ is transient under every deterministic policy.

Proof. We first prove that $I_{\min}(G)$ exists. Every closed set is closable, since $G(s,a) \subseteq I$ for all a implies $G(s,a) \subseteq I$ for some a. Two disjoint nonempty closed sets would therefore have $G \notin \mathcal{M}$. Hence any two nonempty closed sets meet, and their intersection is again closed and nonempty. As $\mathcal{S}$ is finite, the intersection $I_{\min}(G)$ of all nonempty closed sets is closed, nonempty, and the minimum such set.

We next prove that $C(G) \subseteq I_{\min}(G)$. Let I be nonempty closed and suppose $x \in C(G) \setminus I$, say $x \in R \in \mathrm{Rec}(G_\pi)$. Both R and I are closed under G_π, hence so is $R \cap I$. If $R \cap I \neq \varnothing$, minimality of R gives $R \subseteq I$, contradicting $x \notin I$. So $R \cap I = \varnothing$, and (I, R) is a disjoint pair with I closed and R closable, contradicting $G \in \mathcal{M}$.

For the reverse inclusion, write $I := I_{\min}(G)$. Let $\mathrm{cl}_G(y)$ be the smallest closed set containing y, that is, the set of states reachable from y under arbitrary action choices. For $y \in I$ we have $\mathrm{cl}_G(y) \subseteq I$ because I is closed, and $\mathrm{cl}_G(y) \supseteq I$ because $\mathrm{cl}_G(y)$ is nonempty and closed. Hence $\mathrm{cl}_G(y) = I$, so every state of I reaches every state of I.

Fix $y \in I$ and let $d(z)$ be the length of a shortest path from z to y inside I, finite for every $z \in I$ with $d(y) = 0$. For $z \in I \setminus \{y\}$ choose $a_z \in \mathcal{A}$ and $w_z \in G(z, a_z)$ with $d(w_z) = d(z) - 1$; at y choose $a_y \in \mathcal{A}$ arbitrarily. Let π take these actions on I and arbitrary actions elsewhere.

Then y is reachable in G_π from every $z \in I$: for $z \neq y$ by induction on $d(z)$ along $z \to w_z$, and for $z = y$ because $G_\pi(y) = G(y, a_y) \subseteq I$ by closedness, so every $w \in G(y, a_y)$ has $d(w) < \infty$ and reaches y.

Finally I is closed under G, hence under G_π, so it contains some $R \in \mathrm{Rec}(G_\pi)$. Pick $z \in R \subseteq I$. Since y is reachable from z in G_π and R is G_π-closed, $y \in R$, so $y \in C(G)$.

(i) and (ii) follow from $C(G) = I_{\min}(G)$, (iii) from $\mathrm{cl}_G(y) = I$ for $y \in I$, and (iv) from the definition of $C(G)$. □

Lemma 6 also identifies $\mathcal{M}$ with the usual structural description of weakly communicating models. For $G \in \mathfrak{G}$ the following are equivalent: (a) $G \in \mathcal{M}$; (b) $\mathcal{S} = C \sqcup T$ with C closed, all states of C mutually reachable, and every state of T transient under every deterministic policy. Indeed (a) implies (b) with $C = C(G)$. Conversely, assume (b) and let I be closed and J closable with $I \cap J = \varnothing$. Any closable set contains a recurrent class of the policy witnessing its closability, and any nonempty closed set likewise contains a recurrent class, so both I and J meet C. Choosing $x \in I \cap C$ and $z \in J \cap C$, mutual reachability inside C gives a path from x to z, and I is closed, so $z \in I$, contradicting $I \cap J = \varnothing$.

C.3 Self-revealing support under no creation

Lemma 7. *Let $f \in \mathcal{F}_{\mathrm{ne}}$, fix a stationary randomized policy π and an initial state x, and let $R^\pi_P(x)$ be the states reachable from x in the nominal policy graph*

$$\mathcal{G}(P)_\pi(s) := \bigcup_{a:\pi(a|s)>0} \mathcal{G}(P)(s, a).$$

Every trajectory from x under π and every $\widetilde{Q} \in \mathcal{U}^P(\delta)$ remains in $R^\pi_P(x)$. Consequently $g^\pi_{\mathrm{rob},P}(x)$ depends only on the rewards and nominal rows

$$\{P_{s,a} : s \in R^\pi_P(x),\ \pi(a \mid s) > 0\}.$$

In particular, if another center P' agrees with P on those rows and has the same rewards, then $R^\pi_{P'}(x) = R^\pi_P(x)$ and $g^\pi_{\mathrm{rob},P'}(x) = g^\pi_{\mathrm{rob},P}(x)$.

Proof. Because $f \in \mathcal{F}_{\mathrm{ne}}$, Theorem 1 gives $\mathcal{G}(\widetilde{Q})(s, a) \subseteq \mathcal{G}(P)(s, a)$ for every feasible row. The definition of reachability therefore makes $R^\pi_P(x)$ closed under $\widetilde{Q}_\pi$, proving the trajectory claim. Rows based outside this set, and rows of actions assigned zero probability by π, do not enter the law of the trajectory or its average reward. The rectangular infimum can thus be restricted to the displayed rows. If P' agrees on them, the same induction gives the same reachable set and identical rowwise ambiguity sets on every row that can affect the trajectory; taking the two infima proves the last equality. □

C.4 Monotonicity of weak communication

Lemma 1. If $H \subseteq H'$ and $H \in \mathcal{M}$, then $H' \in \mathcal{M}$. Consequently, writing $H_{\geq\theta} := \{e \in \mathcal{G}(P) : \theta_f(P(e)) \geq \theta\}$,

$$\delta^*_f(P) = \max\big\{\theta \in \theta_f(P(\mathcal{G}(P))) : H_{\geq\theta} \in \mathcal{M}\big\}.$$

Proof. For the first claim, suppose $H' \notin \mathcal{M}$ and let I, J be disjoint nonempty sets with I closed and J closable in H'. If $s \in I$ and $a \in \mathcal{A}$ then $H(s, a) \subseteq H'(s, a) \subseteq I$, so I is closed in H; and if $s \in J$ then some a has $H(s, a) \subseteq H'(s, a) \subseteq J$, so J is closable in H. Hence the same pair witnesses $H \notin \mathcal{M}$, and $\mathcal{M}$ is upward closed.

For the second, write $w(H) := \min_{e \in H} \theta_f(P(e))$ and let $\Theta := \theta_f(P(\mathcal{G}(P)))$ be the finite set of edge weights. If $H \subseteq \mathcal{G}(P)$ is nonempty with $H \in \mathcal{M}$, then $\theta := w(H) \in \Theta$ and $H \subseteq H_{\geq\theta}$ by

definition of w, so $H_{\geq\theta} \in \mathcal{M}$ by the first claim, and $w(H_{\geq\theta}) \geq \theta = w(H)$. Every admissible H is therefore dominated by a threshold graph, so the maximum defining $\delta_f^*(P)$ is attained on $\{H_{\geq\theta} : \theta \in \Theta\}$. Conversely each such $H_{\geq\theta}$ lying in $\mathcal{M}$ is admissible with $w(H_{\geq\theta}) \geq \theta$. Taking the maximum over $\theta \in \Theta$ with $H_{\geq\theta} \in \mathcal{M}$ gives the stated identity, both sides being 0 when no such θ exists.

Since $|\Theta| \leq S^2A$, sort Θ. The threshold graphs are nested, and membership in $\mathcal{M}$ is monotone under adding edges, so binary search uses $O(\log(SA))$ membership tests. A candidate H with $H(s,a) = \varnothing$ for some (s,a) is rejected immediately, as it is not a transition graph and hence never belongs to $\mathcal{M}$. Otherwise, for $x \in \mathcal{S}$ let $\mathrm{cl}_H(x)$ be the least fixed point of $B \mapsto B \cup \bigcup_{s\in B, a\in\mathcal{A}} H(s,a)$ started at $\{x\}$, which stabilizes within S passes. If a witness pair (I, J) exists then for any $x \in I$ the set $\mathrm{cl}_H(x)$ is closed, contained in I, and still disjoint from J, so it suffices to test $I = \mathrm{cl}_H(x)$ for the S choices of x. Given such an I, the union of two closable sets is closable, so $\mathcal{S} \setminus I$ has a largest closable subset, namely the greatest fixed point of $B \mapsto \{s \in B : \exists a,\ H(s,a) \subseteq B\}$ started at $\mathcal{S} \setminus I$, again stabilising within S passes. Then $H \notin \mathcal{M}$ if and only if that greatest fixed point is nonempty for some x. Both fixed points can be maintained by a work list with successor counters. Over all x, each state–action–successor incidence is processed $O(S)$ times, so one membership test costs $O(S^3A)$ graph operations. Binary search therefore gives

$$O\big(S^3A\log(SA)\big)$$

operations in total. □

C.5 The common robust core threshold

Theorem 1. For every P, $f \in \mathcal{F}$ and $\delta > 0$,

$$\begin{aligned} \mathcal{G}_\wedge(P) &= \{e \in \mathcal{G}(P) : \theta_f(P(e)) > \delta\}, \\ \mathcal{G}_\wedge(P) \in \mathcal{M} &\iff \delta < \delta_f^*(P), \\ \delta < \delta_f^*(P) &\implies \mathcal{U}^P(\delta) \subseteq \{Q : \mathcal{G}(Q) \in \mathcal{M}\}. \end{aligned}$$

Moreover, quantifying over all instances,

$$\begin{aligned} f \in \mathcal{F}_{\mathrm{nc}} &\iff \big[\mathcal{G}_\wedge(Q) = \mathcal{G}(Q)\ \ \forall Q,\ \forall\delta \in (0,\infty)\big], \\ f \in \mathcal{F}_{\mathrm{ne}} &\iff \big[\mathcal{G}(\widetilde{Q}) \subseteq \mathcal{G}(Q)\ \ \forall Q,\ \forall\delta \in (0,\infty),\ \forall\widetilde{Q} \in \mathcal{U}^Q(\delta)\big], \end{aligned}$$

so $\mathcal{F}_{\mathrm{nc}}$ is exactly the no-deletion class and $\mathcal{F}_{\mathrm{ne}}$ the no-creation class. The quantifiers cannot be moved inside: at a single (Q,δ) the right-hand identities can hold vacuously, for instance when every $\theta_f(Q(e))$ exceeds δ, or when Q has full support.

Proof. We prove the three items in order.

We first identify the robust core. Fix a row $q = P_{s,a}$ and a nominally supported successor i, so $q_i > 0$. By definition, the edge (s,a,i) fails to belong to the robust core exactly when at least one feasible row deletes it:

$$\begin{aligned} i \notin \bigcap_{p\in\mathcal{U}^q(\delta)} \mathrm{supp}(p) &\iff \exists p \in \mathcal{U}^q(\delta) : p_i = 0 \\ &\iff \inf_{\substack{p\in\Delta(\mathcal{S}) \\ p_i=0}} D_f(p\|q) \leq \delta. \end{aligned}$$

Lemma 2, applied with $B = \{i\}$, evaluates the last infimum as $\theta_f(q_i)$. Therefore

$$\begin{aligned} i \notin \bigcap_{p\in\mathcal{U}^q(\delta)} \mathrm{supp}(p) &\iff \theta_f(q_i) \leq \delta, \\ i \in \bigcap_{p\in\mathcal{U}^q(\delta)} \mathrm{supp}(p) &\iff \theta_f(q_i) > \delta. \end{aligned}$$

The ambiguity set is rectangular, so its rows range independently over $\mathcal{U}^{P_{s,a}}(\delta)$. Applying the preceding equivalence to every triple $e = (s,a,s')$ gives

$$\mathcal{G}_\wedge(P) = \{e \in \mathcal{G}(P) : \theta_f(P(e)) > \delta\}.$$

Suppose $f(0) = +\infty$. For every $x \in (0,1)$, the first term in $\theta_f(x) = xf(0) + (1-x)f(1/(1-x))$ is $+\infty$. Hence $\theta_f(x) = +\infty$ for $x \in (0,1)$, and at $x = 1$ the convention $\theta_f(1) = f(0) + f'(\infty) = +\infty$ gives the same value. The deletion criterion therefore shows that no positive-probability edge can be deleted at any finite radius. Conversely, suppose $f(0) < +\infty$. Choose any $x \in (0,1)$ and the two-point row $q = (x, 1-x)$. Since f is finite on $(0,\infty)$,

$$\theta_f(x) = xf(0) + (1-x)f\left(\frac{1}{1-x}\right) < +\infty.$$

At the finite radius $\delta = \theta_f(x)$, Lemma 2 produces a feasible row that deletes the edge of mass x. Thus the property that no supported edge can ever be deleted, uniformly over finite rows and finite radii, holds exactly for $f \in \mathcal{F}_{\mathrm{nc}}$.

Now fix a nominally unsupported successor i, so $q_i = 0$. If $f'(\infty) = +\infty$, Lemma 3 shows that every row with $p_i > 0$ has infinite divergence, and the edge cannot be created at a finite radius. If $f'(\infty) < +\infty$, Lemma 3 constructs, for every $\delta > 0$, a row in $\mathcal{U}^q(\delta)$ with $p_i > 0$. This proves the creation assertion and its converse: the property that no nominally absent edge can ever be created, uniformly over rows and finite radii, holds exactly for $f \in \mathcal{F}_{\mathrm{ne}}$. In both equivalences the quantification over (Q, δ) is essential, since a single instance may satisfy either right-hand condition.

We next derive the threshold for weak communication. For every transition graph $H \subseteq \mathcal{G}(P)$ define

$$w(H) := \min_{e \in H} \theta_f(P(e)).$$

Because H has finitely many edges,

$$\begin{aligned} H \subseteq \mathcal{G}_\wedge(P) &\iff \theta_f(P(e)) > \delta \quad \text{for every } e \in H \\ &\iff w(H) > \delta. \end{aligned}$$

We next verify the monotonicity of weak communication under the addition of edges. Let $H \subseteq G$ and assume $H \in \mathcal{M}$. If, toward a contradiction, $G \notin \mathcal{M}$, then there are disjoint nonempty sets I, J such that I is closed in G and J is closable in G. Closedness of I means

$$G(s,a) \subseteq I, \qquad s \in I, a \in \mathcal{A}.$$

Since $H(s,a) \subseteq G(s,a)$, it follows that $H(s,a) \subseteq I$ for the same states and actions, so I is closed in H. Closability of J means that for every $s \in J$ there is an action a_s with $G(s,a_s) \subseteq J$. Again using $H \subseteq G$, we obtain $H(s,a_s) \subseteq J$, so J is closable in H. This contradicts $H \in \mathcal{M}$. Therefore every supergraph of a weakly communicating graph is weakly communicating.

If $\mathcal{G}_\wedge(P) \in \mathcal{M}$, we may choose $H = \mathcal{G}_\wedge(P)$, and then $H \subseteq \mathcal{G}_\wedge(P)$. Conversely, if there exists $H \in \mathcal{M}$ with $H \subseteq \mathcal{G}_\wedge(P)$, the monotonicity just proved implies $\mathcal{G}_\wedge(P) \in \mathcal{M}$. Combining this fact with the characterization by $w(H)$ gives

$$\begin{aligned} \mathcal{G}_\wedge(P) \in \mathcal{M} &\iff \exists H \subseteq \mathcal{G}(P),\ H \in \mathcal{M} : H \subseteq \mathcal{G}_\wedge(P) \\ &\iff \exists H \subseteq \mathcal{G}(P),\ H \in \mathcal{M} : w(H) > \delta \\ &\iff \delta < \max_{\substack{H \subseteq \mathcal{G}(P) \\ H \in \mathcal{M}}} w(H) \\ &\iff \delta < \delta_f^*(P). \end{aligned}$$

If the maximizing family is empty, the maximum is defined as zero; since $\delta > 0$, both sides of the equivalence are then false.

To compute $\delta_f^*(P)$, list the at most S^2A values $\theta_f(P(e))$ and sort their distinct finite or extended-real values. For each distinct value θ, form the threshold graph of Lemma 1,

$$H_{\geq\theta} := \{e \in \mathcal{G}(P) : \theta_f(P(e)) \geq \theta\},$$

and test whether it is weakly communicating. Such a test is polynomial: compute the maximal end components of the finite state–action graph by repeated strongly connected component decomposition and deletion of state–action pairs whose successors leave the current component; then check whether the states recurrent under some deterministic policy form one communicating class and whether every remaining state is transient under every deterministic policy. Each deletion removes a state

or state–action pair, so there are only polynomially many iterations, and each strongly connected component computation is polynomial. The largest θ whose threshold graph passes this test is exactly the maximum in Lemma 1. Hence $\delta_f^*(P)$ is polynomial-time computable.

It remains to prove universal communication below $\delta_f^*(P)$. By the definition of an intersection, every feasible kernel satisfies

$$\mathcal{G}_\wedge(P) \subseteq \mathcal{G}(\widetilde{Q}), \qquad \widetilde{Q} \in \mathcal{U}^P(\delta).$$

If $\delta < \delta_f^*(P)$, the equivalence proved above gives $\mathcal{G}_\wedge(P) \in \mathcal{M}$. The support graph $\mathcal{G}(\widetilde{Q})$ is a supergraph of $\mathcal{G}_\wedge(P)$, so the edge-addition monotonicity proved above yields $\mathcal{G}(\widetilde{Q}) \in \mathcal{M}$. This holds for every $\widetilde{Q} \in \mathcal{U}^P(\delta)$ and proves the final item. □

Lemma 8. *For a nonempty $B \subseteq \mathcal{S}$, there is a choice of feasible rows that makes B closed if and only if $\rho^P_{f,\mathrm{all}}(B) \le \delta$. There is a deterministic policy and a choice of feasible rows that makes B closed under that policy if and only if $\rho^P_{f,\mathrm{pol}}(B) \le \delta$.*

Proof. Fix $s \in B$ and $a \in \mathcal{A}$. A candidate row $p \in \Delta(\mathcal{S})$ has support contained in B if and only if it assigns zero mass to B^c:

$$\mathrm{supp}(p) \subseteq B \iff p(B^c) = 0.$$

Lemma 2, applied to the nominal row $P_{s,a}$ and the set B^c, gives

$$\begin{aligned} &\exists p \in \mathcal{U}^{P_{s,a}}(\delta) : \mathrm{supp}(p) \subseteq B \\ &\quad\iff \inf_{\substack{p \in \Delta(\mathcal{S}) \\ p(B^c)=0}} D_f(p \| P_{s,a}) \le \delta \\ &\quad\iff \theta_f\big(P(B^c \mid s,a)\big) \le \delta. \end{aligned}$$

To make B closed, a feasible row supported on B must be chosen for every pair (s,a) with $s \in B$ and $a \in \mathcal{A}$. Because the ambiguity set is rectangular, these choices impose no cross-row compatibility condition. Thus such a simultaneous choice exists if and only if

$$\theta_f\big(P(B^c \mid s,a)\big) \le \delta \quad \text{for every } s \in B, a \in \mathcal{A}.$$

The last condition is equivalent to

$$\max_{\substack{s \in B \\ a \in \mathcal{A}}} \theta_f\big(P(B^c \mid s,a)\big) = \rho^P_{f,\mathrm{all}}(B) \le \delta,$$

which proves the first assertion.

To make B closed under some deterministic policy, for each state $s \in B$ it is enough and necessary to choose one action a_s whose feasible row can be supported on B. By the rowwise equivalence above, such an action exists exactly when

$$\min_{a \in \mathcal{A}} \theta_f\big(P(B^c \mid s,a)\big) \le \delta. \tag{12}$$

The action set is finite, so the minimum is attained. Choosing an attaining action independently at each state defines a deterministic policy. Requiring inequality (12) for every $s \in B$ is equivalent to

$$\max_{s \in B} \min_{a \in \mathcal{A}} \theta_f\big(P(B^c \mid s,a)\big) = \rho^P_{f,\mathrm{pol}}(B) \le \delta.$$

This proves the second assertion in both directions. □

C.6 The universal weak-communication threshold

Theorem 2. For every P, $f \in \mathcal{F}$ and $\delta > 0$,

$$\begin{aligned} \mathcal{U}^P(\delta) \not\subseteq \{Q : \mathcal{G}(Q) \in \mathcal{M}\} &\iff \exists\, I, J \neq \varnothing,\ I \cap J = \varnothing, \\ &\qquad \rho^P_{f,\mathrm{all}}(I) \le \delta,\ \rho^P_{f,\mathrm{pol}}(J) \le \delta \\ &\iff \delta \ge \delta_f^\dagger(P). \end{aligned}$$

Proof. We first prove the forward implication. Suppose that $\widetilde{Q} \in \mathcal{U}^P(\delta)$ and that its support graph is not weakly communicating. Negating the definition of weak communication produces disjoint nonempty sets $I, J \subseteq \mathcal{S}$ such that

$$\begin{aligned} \operatorname{supp}(\widetilde{Q}_{s,a}) &\subseteq I, && s \in I,\ a \in \mathcal{A}, \\ \operatorname{supp}(\widetilde{Q}_{s,a_s}) &\subseteq J, && s \in J, \end{aligned}$$

for some action $a_s \in \mathcal{A}$ chosen at each $s \in J$. Since every row of $\widetilde{Q}$ is feasible,

$$D_f(\widetilde{Q}_{s,a} \| P_{s,a}) \le \delta \quad \text{for every } (s,a).$$

The first collection of rows therefore realizes I as a closed set, and the second collection realizes J as closed under the deterministic policy $s \mapsto a_s$. Lemma 8 gives

$$\rho^P_{f,\mathrm{all}}(I) \le \delta, \qquad \rho^P_{f,\mathrm{pol}}(J) \le \delta.$$

Thus I is δ-closed and J is δ-closable.

For the reverse implication, suppose that disjoint nonempty I, J satisfy

$$\rho^P_{f,\mathrm{all}}(I) \le \delta, \qquad \rho^P_{f,\mathrm{pol}}(J) \le \delta. \tag{13}$$

For every $s \in I$ and every $a \in \mathcal{A}$, the first inequality in (13) and Lemma 8 provide a row

$$p^I_{s,a} \in \mathcal{U}^{P_{s,a}}(\delta) \quad \text{with} \quad \operatorname{supp}(p^I_{s,a}) \subseteq I.$$

For every $s \in J$, choose an action a_s attaining

$$\min_{a \in \mathcal{A}} \theta_f\big(P(J^c \mid s,a)\big).$$

The second inequality in (13) implies that this minimum is at most δ, so Lemma 8 provides

$$p^J_{s,a_s} \in \mathcal{U}^{P_{s,a_s}}(\delta) \quad \text{with} \quad \operatorname{supp}(p^J_{s,a_s}) \subseteq J.$$

Because $I \cap J = \varnothing$, no row has been assigned two conflicting values. Define $\widetilde{Q}$ using the rows $p^I_{s,a}$ on states in I, the rows p^J_{s,a_s} on the selected state–action pairs in J, and the nominal row $P_{s,a}$ on every remaining state–action pair. The nominal row is feasible because $D_f(P_{s,a} \| P_{s,a}) = 0 \le \delta$. Hence $\widetilde{Q} \in \mathcal{U}^P(\delta)$. By construction, I is closed in $\mathcal{G}(\widetilde{Q})$ and J is closed under the deterministic choices a_s, so J is closable. Therefore $\mathcal{G}(\widetilde{Q})$ is not weakly communicating.

Finally, for each disjoint nonempty pair (I, J) define

$$d(I,J) := \max\left\{\rho^P_{f,\mathrm{all}}(I), \rho^P_{f,\mathrm{pol}}(J)\right\}.$$

The preceding equivalence shows that a non-weakly-communicating feasible kernel exists if and only if there is a pair with $d(I,J) \le \delta$. The state space is finite, so the family of such pairs is finite and its minimum is attained. Therefore

$$\begin{aligned} \exists \widetilde{Q} \in \mathcal{U}^P(\delta) : \mathcal{G}(\widetilde{Q}) \notin \mathcal{M} &\iff \min_{I,J} d(I,J) \le \delta \\ &\iff \delta^\dagger_f(P) \le \delta. \end{aligned}$$

Negating both sides gives

$$\forall \widetilde{Q} \in \mathcal{U}^P(\delta),\ \mathcal{G}(\widetilde{Q}) \in \mathcal{M} \iff \delta < \delta^\dagger_f(P),$$

where the strict inequality is necessary because the closability conditions themselves use $\le \delta$. □

C.7 Threshold comparison and sharp examples

Proposition 2.

$$\begin{aligned} &(\forall Q)(\forall f \in \mathcal{F}) && \delta^*_f(Q) \le \delta^\dagger_f(Q), \\ &(\exists f \in \mathcal{F})(\exists Q : \mathcal{G}(Q) \in \mathcal{M}) && \delta^*_f(Q) = \delta^\dagger_f(Q) < \infty, \\ &(\exists f \in \mathcal{F} \setminus \mathcal{F}_{\mathrm{nc}})(\exists Q' : \mathcal{G}(Q') \in \mathcal{M}) && \delta^*_f(Q') < \delta^\dagger_f(Q'). \end{aligned}$$

Proof. Fix Q and f. For every radius satisfying $0 < \delta < \delta_f^*(Q)$, Theorem 1 implies

$$\mathcal{G}(\widetilde{Q}) \in \mathcal{M} \qquad \text{for every } \widetilde{Q} \in \mathcal{U}^Q(\delta).$$

Theorem 2 says that the universal communication property in this display is equivalent to $\delta < \delta_f^\dagger(Q)$. Hence

$$0 < \delta < \delta_f^*(Q) \quad \Longrightarrow \quad \delta < \delta_f^\dagger(Q).$$

If $\delta_f^*(Q) > \delta_f^\dagger(Q)$, then $\delta_f^\dagger(Q)$ is finite and we may choose

$$\delta \in \big(\delta_f^\dagger(Q), \delta_f^*(Q)\big).$$

The implication above would yield simultaneously $\delta < \delta_f^\dagger(Q)$ and $\delta > \delta_f^\dagger(Q)$, a contradiction. Therefore

$$\delta_f^*(Q) \le \delta_f^\dagger(Q).$$

We first construct a finite equality example. Take the KL generator $f(x) = x \log x$, one action, two states, and $\epsilon \in (0, 1/2)$. Define

$$Q^= = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}.$$

The chain is irreducible, so its support graph is weakly communicating. For KL,

$$\theta_f(x) = -\log(1-x).$$

Every weakly communicating one-action subgraph must contain at least one of the two cross edges, each of nominal mass ϵ. Conversely, the two cross edges alone form an irreducible subgraph. Since $\epsilon < 1 - \epsilon$ and θ_f is increasing,

$$\delta_f^*(Q^=) = \theta_f(\epsilon).$$

For $\delta_f^\dagger$, the only ordered pairs of disjoint nonempty sets are

$$(I, J) = (\{1\}, \{2\}) \quad \text{and} \quad (I, J) = (\{2\}, \{1\}).$$

For either pair, the probability of leaving I and the probability of leaving J are both ϵ. Therefore

$$\rho_{f,\mathrm{all}}^{Q^=}(I) = \rho_{f,\mathrm{pol}}^{Q^=}(J) = \theta_f(\epsilon),$$

and hence

$$0 < \delta_f^*(Q^=) = \delta_f^\dagger(Q^=) = -\log(1-\epsilon) < +\infty.$$

We now construct an instance with strict inequality. Take the KL generator $f(x) = x \log x$, one action, three states, and $\epsilon \in (0, 1/3)$. Define

$$Q'(j \mid i) = \begin{cases} 1 - 2\epsilon, & j = i, \\ \epsilon, & j \neq i. \end{cases}$$

Every entry of Q' is positive, so its one-action Markov chain is irreducible and its support graph is weakly communicating. For KL, $f(0) = 0$ and

$$\begin{aligned} \theta_f(x) &= x f(0) + (1-x) f\left(\frac{1}{1-x}\right) \\ &= (1-x)\frac{1}{1-x}\log\left(\frac{1}{1-x}\right) \\ &= -\log(1-x). \end{aligned}$$

This function is strictly increasing on $[0, 1)$. Each off-diagonal edge has mass ϵ, whereas each self-loop has mass $1 - 2\epsilon$. Because $\epsilon < 1/3$,

$$\epsilon < 1 - 2\epsilon,$$

and hence

$$\theta_f(\epsilon) < \theta_f(1 - 2\epsilon).$$

If $\delta < \theta_f(\epsilon)$, every edge has weight greater than δ, so the robust core is the full irreducible graph. If

$$\theta_f(\epsilon) \le \delta < \theta_f(1-2\epsilon),$$

all six off-diagonal edges are deleted from the robust core while the three self-loops remain. The resulting graph has three disjoint closed classes and is not weakly communicating. Increasing the radius further cannot restore an edge because the robust core is decreasing in δ. Consequently,

$$\delta_f^*(Q') = \theta_f(\epsilon) = -\log(1-\epsilon).$$

It remains to compute $\delta_f^\dagger(Q')$. With one action, the maximum over all actions and the minimum over actions are identical, so

$$\rho_{f,\mathrm{all}}^{Q'}(B) = \rho_{f,\mathrm{pol}}^{Q'}(B) =: \rho_f^{Q'}(B).$$

If $B = \{i\}$ is a singleton, the probability of leaving B from its only state is

$$Q'(B^c \mid i) = 2\epsilon,$$

and therefore

$$\rho_f^{Q'}(\{i\}) = \theta_f(2\epsilon).$$

If $B = \{i, j\}$ has two states, then from either state the probability of moving to the unique state outside B is ϵ. Thus

$$\rho_f^{Q'}(\{i,j\}) = \theta_f(\epsilon).$$

Every disjoint nonempty pair (I, J) in a three-state space contains at least one singleton. Therefore

$$\max\{\rho_f^{Q'}(I), \rho_f^{Q'}(J)\} \ge \theta_f(2\epsilon).$$

Taking I to be a singleton and $J = I^c$ gives equality, since $\theta_f(\epsilon) < \theta_f(2\epsilon)$. Hence

$$\delta_f^\dagger(Q') = \theta_f(2\epsilon) = -\log(1-2\epsilon).$$

Finally, $2\epsilon > \epsilon$ and θ_f is strictly increasing, so

$$\delta_f^\dagger(Q') = -\log(1-2\epsilon) > -\log(1-\epsilon) = \delta_f^*(Q'),$$

which proves strictness. □

C.8 Margin calculus

Lemma 9. *Let $f \in \mathcal{F}$, $\delta > 0$ and $\delta < \delta_f^*(P)$.*

(i) $\epsilon_0(P) = q_\star$ when $C(P) \neq \mathcal{S}$, and $\epsilon_e(P) = P(e) - q_\star$ for every core edge e; every core edge satisfies $P(e) > q_\star$, so $\epsilon_{\mathrm{top}}(P) > 0$ whenever $C(P) = \mathcal{S}$.

(ii) There are $\eta_f \in (0,1)$ and $L_f \in [1,\infty)$, depending only on (f, δ), such that whenever $C(P) = \mathcal{S}$,

$$\epsilon_{\mathrm{top}}(P) \;\ge\; \min\Big\{\eta_f,\; \frac{\Delta_\theta(P)}{L_f},\; p_{\min,C}(P)\Big\}.$$

(iii) If $C(P) = \mathcal{S}$ and $\|Q-P\|_{\max,1} < 2\epsilon_{\mathrm{top}}(P)$ then $\mathcal{G}_\wedge(P) \subseteq \mathcal{G}_\wedge(Q)$ and hence $\delta < \delta_f^*(Q) \le \delta_f^\dagger(Q)$; consequently $\epsilon_{\mathrm{wc}}(P_R) \ge 2\epsilon_{\mathrm{top}}(P)$ and $\bar{\epsilon}_{\mathrm{wc}}(P_R) \ge \min\{\frac{1}{2}, \epsilon_{\mathrm{top}}(P), \frac{1}{2}p_{\min,R}(P)\}$.

(iv) If $f \in \mathcal{F}_{\mathrm{ne}}$, then $\delta < \delta_f^*(P_R)$ and

$$\Delta_{\mathrm{bell}}(P_R) \;\ge\; \min\Big\{\eta_f,\; \frac{\Delta_\theta(P)}{L_f},\; p_{\min,C}(P)\Big\}.$$

Proof. Throughout write $q_\star = q_\star(f,\delta)$ and $A := \{y \in [0,1] : \theta_f(y) > \delta\}$. By Lemma 5,

$$(q_\star, 1] \subseteq A \subseteq [q_\star, 1]. \tag{14}$$

Since $\delta < \delta_f^*(P)$, Theorem 1 gives $\mathcal{G}_\wedge(P) \in \mathcal{M}$, so $\mathcal{G}_\wedge(P)$ has nonempty rows and in particular contains some edge e with $\theta_f(P(e)) > \delta$. Monotonicity (Lemma 4) then gives $\theta_f(1) \geq \theta_f(P(e)) > \delta$, so $q_\star < 1$ by Lemma 5.

(i). Suppose $C(P) \neq \mathcal{S}$. Monotonicity of θ_f gives $\sup_{x\in[0,\epsilon]} \theta_f(x) = \theta_f(\epsilon)$, so the set of admissible radii in the definition of $\epsilon_0(P)$ is $\{\epsilon > 0 : \theta_f(\epsilon) \leq \delta\}$, whose supremum is exactly $q_\star$ by definition of $q_\star$.

For an edge e, monotonicity gives $\inf_{x\in I_e(\epsilon)} \theta_f(x) = \theta_f\big((P(e) - \epsilon)_+\big)$, so the admissible radii are $\{\epsilon > 0 : (P(e) - \epsilon)_+ \in A\}$. If $0 < \epsilon < P(e) - q_\star$ then $P(e) - \epsilon > q_\star$, which lies in A by (14); if $\epsilon > P(e) - q_\star$ then $(P(e) - \epsilon)_+ < q_\star$, which lies outside A by (14). Hence $\epsilon_e(P) = P(e) - q_\star$.

That $P(e) > q_\star$ strictly for a core edge is seen as follows. Theorem 1 gives $P(e) \in A$, so $P(e) \geq q_\star$ by (14). If $f(0) = +\infty$ then $q_\star = 0$ while $P(e) > 0$, since $e \in \mathcal{G}(P)$. If $f(0) < +\infty$ then, because $q_\star < 1$, Lemma 5 gives $\theta_f(q_\star) = \delta$, so $P(e) = q_\star$ would contradict $\theta_f(P(e)) > \delta$. Finally, when $C(P) = \mathcal{S}$ the convention $\epsilon_0(P) = +\infty$ applies and $\epsilon_{\mathrm{top}}(P) = \min\{P(e) - q_\star : e \in \mathcal{G}_\wedge(P)\} > 0$, the minimum being over the finite nonempty edge set of $\mathcal{G}_\wedge(P)$.

(ii). Assume $C(P) = \mathcal{S}$. If $f(0) = +\infty$, then $q_\star = 0$ and part (i) gives $\epsilon_{\mathrm{top}}(P) = \min_{e\in\mathcal{G}_\wedge(P)} P(e) \geq p_{\min,C}(P)$, since every edge of $\mathcal{G}_\wedge(P)$ is a positive nominal transition based in $C(P) = \mathcal{S}$. Setting $\eta_f := \frac{1}{2}$ and $L_f := 1$ in this case gives the claim, because $\Delta_\theta(P) = +\infty$ there.

If $f(0) < +\infty$, let $\eta_f := \frac{1}{2}(1 - q_\star) \in (0, 1)$, so that $[q_\star, q_\star + \eta_f]$ is a compact subinterval of $[0, 1)$. By Lemma 4 the map θ_f is convex and finite on $[0, 1)$, hence Lipschitz on that compact subinterval; let $L_f \geq 1$ be a Lipschitz constant there. Both η_f and L_f depend only on (f, δ). Now fix a core edge e. If $P(e) - q_\star \geq \eta_f$ then $\epsilon_e(P) \geq \eta_f$ by part (i). Otherwise $P(e) \in (q_\star, q_\star + \eta_f)$ and, using $\theta_f(q_\star) = \delta$ from Lemma 5,

$$\Delta_e := \theta_f(P(e)) - \delta = \theta_f(P(e)) - \theta_f(q_\star) \leq L_f\big(P(e) - q_\star\big) = L_f\,\epsilon_e(P),$$

so $\epsilon_e(P) \geq \Delta_e/L_f \geq \Delta_\theta(P)/L_f$. Taking the minimum over the finitely many core edges gives $\epsilon_{\mathrm{top}}(P) \geq \min\{\eta_f, \Delta_\theta(P)/L_f\}$, which implies the stated bound.

(iii). Assume $C(P) = \mathcal{S}$, so $P_R = P$, and let $\|Q - P\|_{\max,1} < 2\epsilon_{\mathrm{top}}(P)$. For any (s, a) and any $B \subseteq \mathcal{S}$ one has $|Q(B \mid s, a) - P(B \mid s, a)| \leq \frac{1}{2}\|Q_{s,a} - P_{s,a}\|_1$, because the two signed masses $Q(B) - P(B)$ and $Q(B^c) - P(B^c)$ are opposite and their absolute values sum to at most $\|Q_{s,a} - P_{s,a}\|_1$. Applying this with $B = \{s'\}$ to a core edge $e = (s, a, s')$ and using part (i),

$$|Q(e) - P(e)| < \epsilon_{\mathrm{top}}(P) \leq \epsilon_e(P) = P(e) - q_\star,$$

so $Q(e) > q_\star \geq 0$. In particular $Q(e) > 0$, so $e \in \mathcal{G}(Q)$, and $\theta_f(Q(e)) > \delta$ by (14). Theorem 1 therefore places e in $\mathcal{G}_\wedge(Q)$, and we conclude $\mathcal{G}_\wedge(P) \subseteq \mathcal{G}_\wedge(Q)$.

Since $\delta < \delta_f^*(P)$, Theorem 1 gives $\mathcal{G}_\wedge(P) \in \mathcal{M}$; Lemma 1 then gives $\mathcal{G}_\wedge(Q) \in \mathcal{M}$, and Theorem 1 once more gives $\delta < \delta_f^*(Q)$. Proposition 2 gives $\delta_f^*(Q) \leq \delta_f^\dagger(Q)$, so $\delta < \delta_f^\dagger(Q)$.

Every radius $\epsilon < 2\epsilon_{\mathrm{top}}(P)$ is therefore admissible in the definition of $\epsilon_{\mathrm{wc}}(P_R)$, whence $\epsilon_{\mathrm{wc}}(P_R) \geq 2\epsilon_{\mathrm{top}}(P)$ and

$$\begin{aligned}\bar{\epsilon}_{\mathrm{wc}}(P_R) &= \tfrac{1}{2}\min\{1, \epsilon_{\mathrm{wc}}(P_R), p_{\min,R}(P)\} \\ &\geq \tfrac{1}{2}\min\{1, 2\epsilon_{\mathrm{top}}(P), p_{\min,R}(P)\} \\ &= \min\{\tfrac{1}{2}, \epsilon_{\mathrm{top}}(P), \tfrac{1}{2}p_{\min,R}(P)\}.\end{aligned}$$

(iv). Let $f \in \mathcal{F}_{\mathrm{ne}}$. Lemma 18 gives $\delta < \delta_f^*(P_R)$, and $C(P_R) = R(P)$ is the full state space of the restricted kernel. Parts (i)–(ii), applied to P_R, therefore give

$$\epsilon_{\mathrm{top}}(P_R) \geq \min\Big\{\eta_f, \frac{\Delta_\theta(P_R)}{L_f}, p_{\min,R}(P)\Big\}.$$

Restriction preserves every recurrent core edge and its deletion cost, so $\Delta_\theta(P_R) = \Delta_\theta(P)$ and $p_{\min,R}(P) = p_{\min,C}(P)$. For any $(I, J) \in \mathfrak{W}_{\mathrm{bell}}(P_R)$, the set J is closable in $\mathcal{G}_\wedge(P_R)$: choose the nominally internal action at each state of J; its core row is nonempty and remains supported in J. If no core edge left I, then I would be closed under every action. It would therefore contain the

communicating class of $\mathcal{G}_\wedge(P_R)$, whereas the policy closing J would give a recurrent class inside the disjoint set J. This contradicts weak communication. Hence some core edge $e = (s, a, s')$ has $s \in I$ and $s' \notin I$, and part (i) yields

$$P_R(I^c \mid s, a) - q_\star \geq P_R(e) - q_\star \geq \epsilon_{\mathrm{top}}(P_R).$$

Thus $m_{\mathrm{bell}}(P_R) \geq \epsilon_{\mathrm{top}}(P_R)$; combining this with the definition of Δ_{bell} proves (iv). □

C.9 Robust Bellman equation and vanishing discount

Theorem 3. For fixed P, $f \in \mathcal{F}$, and $\delta > 0$, the following are equivalent: (i) $\delta < \delta_f^{\mathrm{bell}}(P)$; (ii) every fixed point of the recession operator $\widehat{\mathcal{L}}_P$ is constant; and (iii) for every $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, $\mathrm{Eig}(\mathcal{L}_P^{\mathrm{rob}})$ is nonempty, all its eigenpairs share the same gain $g^{\mathrm{rob}}(P)$, and every selector π^* greedy for any such bias satisfies $g^{\pi^*}_{\mathrm{rob},P}(s) = g^*_{\mathrm{rob},P}(s) = g^{\mathrm{rob}}(P)$ for every $s \in \mathcal{S}$. Under these conditions the normalized discounted biases are bounded, their accumulation points are robust biases, and $(1-\gamma)v_\gamma \to g^{\mathrm{rob}}(P)\mathbf{1}$. If $\delta \geq \delta_f^{\mathrm{bell}}(P)$, some reward has state-dependent robust optimal gain and no scalar additive eigenpair. If $f \notin \mathcal{F}_{\mathrm{ne}}$, then $\delta_f^{\mathrm{bell}}(P) = +\infty$.

Proof. We first record the elementary properties of the row functional. For every nominal row q, the set $\mathcal{U}^q(\delta)$ is nonempty because it contains q. It is compact because it is a lower-level set of the lower-semicontinuous function $p \mapsto D_f(p\|q)$ inside the compact simplex. Hence the minimum defining $\sigma_\delta^f(q,u)$ is attained.

For any $u, v \in \mathbb{R}^{\mathcal{S}}$ and $c \in \mathbb{R}$, the row functional satisfies

$$\begin{aligned} u \leq v &\Longrightarrow \sigma_\delta^f(q,u) \leq \sigma_\delta^f(q,v), \\ \sigma_\delta^f(q, u + c\mathbf{1}) &= \sigma_\delta^f(q,u) + c, \\ \sigma_\delta^f(q,\alpha u) &= \alpha\sigma_\delta^f(q,u), \qquad \alpha \geq 0, \\ |\sigma_\delta^f(q,u) - \sigma_\delta^f(q,v)| &\leq \|u - v\|_\infty. \end{aligned} \tag{15}$$

The first three identities follow directly by applying the corresponding operation inside $p^\top u$ and using $p^\top \mathbf{1} = 1$. For the last one, let p_v minimize $p^\top v$. Then

$$\begin{aligned} \sigma_\delta^f(q,u) - \sigma_\delta^f(q,v) &\leq p_v^\top u - p_v^\top v \\ &= p_v^\top (u - v) \\ &\leq \|u - v\|_\infty. \end{aligned}$$

Interchanging u and v proves the absolute-value bound.

Define the recession operator

$$[\widehat{\mathcal{L}}_P u](s) := \max_{a\in\mathcal{A}} \sigma_\delta^f(P_{s,a}, u).$$

The properties in equations (15), together with the fact that the maximum is over finitely many actions, show that $\widehat{\mathcal{L}}_P$ is monotone, additively homogeneous, positively homogeneous, and continuous.

We next prove that $\delta < \delta_f^{\mathrm{bell}}(P)$ forces every recession fixed point to be constant. Suppose that

$$u = \widehat{\mathcal{L}}_P u.$$

Set

$$m := \min_s u(s), \qquad M := \max_s u(s),$$

and define the nonempty level sets

$$I := \{s : u(s) = m\}, \qquad J := \{s : u(s) = M\}.$$

Assume for contradiction that u is nonconstant. Then $m < M$ and $I \cap J = \varnothing$.

Fix $s \in I$. For every action a and every probability vector p,

$$p^\top u = \sum_{s'} p(s')u(s') \geq \sum_{s'} p(s')m = m.$$

Therefore

$$\sigma_\delta^f(P_{s,a}, u) \geq m \qquad \text{for every } a.$$

On the other hand, the fixed-point identity at s gives

$$m = u(s) = \max_{a \in \mathcal{A}} \sigma_\delta^f(P_{s,a}, u).$$

The maximum of numbers that are all at least m can equal m only if every one of them equals m. Hence

$$\sigma_\delta^f(P_{s,a}, u) = m \qquad \text{for every } a.$$

Let $p_{s,a}$ be a minimizing row. Then

$$0 = p_{s,a}^\top u - m = \sum_{s'} p_{s,a}(s')\big(u(s') - m\big).$$

Every summand is nonnegative. Their sum can be zero only if $p_{s,a}(s') = 0$ whenever $u(s') > m$. Thus

$$\operatorname{supp}(p_{s,a}) \subseteq I \qquad (s \in I,\ a \in \mathcal{A}).$$

All these rows are feasible, so $\rho^P_{f,\mathrm{all}}(I) \leq \delta$.

Now fix $s \in J$. The fixed-point identity gives

$$M = u(s) = \max_{a \in \mathcal{A}} \sigma_\delta^f(P_{s,a}, u).$$

Because the action set is finite, some action a_s attains this maximum:

$$\sigma_\delta^f(P_{s,a_s}, u) = M.$$

Every probability vector satisfies $p^\top u \leq M$. Since the minimum of $p^\top u$ over the feasible set equals M, every feasible row must in fact have expectation M. For any such row,

$$0 = M - p^\top u = \sum_{s'} p(s')\big(M - u(s')\big),$$

and again every summand is nonnegative. Hence every feasible row for (s, a_s) is supported on J.

If $f \notin \mathcal{F}_{\mathrm{ne}}$, choose any $i \notin J$. If $P_{s,a_s}(i) > 0$, the feasible nominal row contradicts the preceding support conclusion. If $P_{s,a_s}(i) = 0$, the creation lemma constructs, at every positive radius, a feasible row assigning positive mass to i, again a contradiction. Thus a nonconstant fixed point is impossible when creation is possible.

Suppose instead that $f \in \mathcal{F}_{\mathrm{ne}}$. The nominal row is feasible, so $\operatorname{supp}(P_{s,a_s}) \subseteq J$ for every $s \in J$; hence J is closable in $\mathcal{G}(P)$. The disjoint pair (I, J) is admissible in equation (4), and

$$\delta_f^{\mathrm{bell}}(P) \leq \rho^P_{f,\mathrm{all}}(I) \leq \delta,$$

contradicting $\delta < \delta_f^{\mathrm{bell}}(P)$. Thus every fixed point of $\widehat{\mathcal{L}}_P$ is constant under the claimed condition.

We now construct the discounted fixed points and prove that their normalized biases are bounded. For $\gamma \in (0, 1)$ and $u, v \in \mathbb{R}^{\mathcal{S}}$,

$$\begin{aligned}
&\big|[\mathcal{L}^{\mathrm{rob}}_{P,\gamma} u](s) - [\mathcal{L}^{\mathrm{rob}}_{P,\gamma} v](s)\big| \\
&\quad \leq \gamma \max_{a \in \mathcal{A}} |\sigma_\delta^f(P_{s,a}, u) - \sigma_\delta^f(P_{s,a}, v)| \\
&\quad \leq \gamma \|u - v\|_\infty.
\end{aligned}$$

Taking the maximum over s shows that $\mathcal{L}^{\mathrm{rob}}_{P,\gamma}$ is a γ-contraction in the supremum norm. Banach's fixed-point theorem therefore gives a unique fixed point v_γ.

The constant vector 0 is a subsolution because rewards are nonnegative:

$$\mathcal{L}^{\mathrm{rob}}_{P,\gamma} 0 \geq 0.$$

The constant vector $c_\gamma \mathbf{1}$, where $c_\gamma := 1/(1-\gamma)$, is a supersolution because

$$\begin{aligned}[\mathcal{L}^{\mathrm{rob}}_{P,\gamma}(c_\gamma \mathbf{1})](s) &= \max_a \{r(s,a) + \gamma c_\gamma\} \\ &\le 1 + \gamma c_\gamma = c_\gamma.\end{aligned}$$

Monotonicity and convergence of value iteration to the unique fixed point therefore give

$$0 \le v_\gamma \le \frac{1}{1-\gamma}\mathbf{1}. \tag{16}$$

Fix $s_\circ \in \mathcal{S}$ and define

$$h_\gamma := v_\gamma - v_\gamma(s_\circ)\mathbf{1}, \qquad g_\gamma := (1-\gamma)v_\gamma(s_\circ).$$

Inequality (16) implies $g_\gamma \in [0,1]$. Using additive homogeneity of the row functional, the discounted fixed-point equation becomes

$$\begin{aligned}[\mathcal{L}^{\mathrm{rob}}_{P,\gamma} h_\gamma](s) &= \max_a \Big\{ r(s,a) + \gamma\big(\sigma^f_\delta(P_{s,a}, v_\gamma) - v_\gamma(s_\circ)\big)\Big\} \\ &= v_\gamma(s) - \gamma v_\gamma(s_\circ) \\ &= h_\gamma(s) + (1-\gamma)v_\gamma(s_\circ).\end{aligned}$$

Thus

$$g_\gamma \mathbf{1} + h_\gamma = \mathcal{L}^{\mathrm{rob}}_{P,\gamma} h_\gamma. \tag{17}$$

We now prove that the family $\{h_\gamma : \gamma \in (0,1)\}$ is bounded. For every fixed $\bar\gamma < 1$ and every $\gamma \le \bar\gamma$,

$$\|h_\gamma\|_\infty \le \mathrm{sp}(v_\gamma) \le \frac{1}{1-\bar\gamma},$$

so it remains to treat the regime $\gamma \uparrow 1$. Suppose the family were unbounded there. Since $h_\gamma(s_\circ) = 0$, unboundedness in supremum norm is equivalent to unbounded span. Hence there would be a sequence $\gamma_n \uparrow 1$ such that

$$L_n := \mathrm{sp}(h_{\gamma_n}) \longrightarrow +\infty.$$

Define

$$y_n := \frac{h_{\gamma_n}}{L_n}.$$

Then $y_n(s_\circ) = 0$ and $\mathrm{sp}(y_n) = 1$. These two facts imply $\|y_n\|_\infty \le 1$, so compactness of the finite-dimensional cube $[-1,1]^{\mathcal{S}}$ gives a convergent subsequence. Relabel it so that $y_n \to y$. Continuity of the maximum and minimum over the finite state space gives

$$y(s_\circ) = 0, \qquad \mathrm{sp}(y) = 1.$$

Divide equation (17) by L_n. Positive homogeneity of σ^f_δ gives, for every state s,

$$\frac{g_{\gamma_n}}{L_n} + y_n(s) = \max_a \left\{ \frac{r(s,a)}{L_n} + \gamma_n \sigma^f_\delta(P_{s,a}, y_n) \right\}.$$

Now $g_{\gamma_n}/L_n \to 0$, $r(s,a)/L_n \to 0$, and $\gamma_n \to 1$. The Lipschitz property in equations (15) yield

$$\sigma^f_\delta(P_{s,a}, y_n) \to \sigma^f_\delta(P_{s,a}, y)$$

for every state–action pair. Since there are finitely many actions, we may pass the limit through the maximum and obtain

$$y(s) = \max_a \sigma^f_\delta(P_{s,a}, y) = [\widehat{\mathcal{L}}_P y](s).$$

Thus y is a fixed point of the recession operator. The preceding fixed-point argument shows that y must be constant, whereas $\mathrm{sp}(y) = 1$. This contradiction proves that the normalized discounted biases are bounded.

We next pass to the average-reward equation. Let $\gamma_n \uparrow 1$ be arbitrary. The scalars g_{γ_n} lie in $[0,1]$, and the vectors h_{γ_n} lie in a fixed compact subset of $\mathbb{R}^{\mathcal{S}}$. Hence there is a subsequence, not relabeled, and a pair (g,h) such that

$$g_{\gamma_n} \to g, \qquad h_{\gamma_n} \to h.$$

For each state–action pair, the row Lipschitz bound gives

$$\sigma_\delta^f(P_{s,a}, h_{\gamma_n}) \to \sigma_\delta^f(P_{s,a}, h).$$

Taking limits in equation (17) and using $\gamma_n \to 1$ yields

$$g\mathbf{1} + h = \mathcal{L}_P^{\mathrm{rob}} h.$$

The additive eigenvalue is unique. To see this, suppose

$$g\mathbf{1} + h = \mathcal{L}_P^{\mathrm{rob}} h, \qquad g'\mathbf{1} + h' = \mathcal{L}_P^{\mathrm{rob}} h'.$$

Let

$$d := \max_s \{h(s) - h'(s)\}$$

and choose s^* attaining the maximum. Then $h \le h' + d\mathbf{1}$. Monotonicity and additive homogeneity imply

$$\mathcal{L}_P^{\mathrm{rob}} h \le \mathcal{L}_P^{\mathrm{rob}}(h' + d\mathbf{1}) = \mathcal{L}_P^{\mathrm{rob}} h' + d\mathbf{1}.$$

Evaluating at s^* gives

$$\begin{aligned} g + h(s^*) &\le g' + h'(s^*) + d \\ &= g' + h(s^*), \end{aligned}$$

and hence $g \le g'$. Interchanging the two eigenpairs gives $g' \le g$, so $g = g'$.

Every convergent subsequence of g_γ must therefore have the same limit g. Since the family is bounded, this implies the full convergence $g_\gamma \to g$. Moreover,

$$(1-\gamma)v_\gamma = g_\gamma \mathbf{1} + (1-\gamma)h_\gamma.$$

The second term converges to zero because h_γ is uniformly bounded. Therefore

$$(1-\gamma)v_\gamma \longrightarrow g\mathbf{1}.$$

For any convergent subsequence of h_γ, taking limits in equation (17), as above, shows that its limit is a robust bias.

It remains to verify robust optimality. Let π^* be any deterministic selector greedy with respect to h. The Bellman equation gives, for every state,

$$g + h(s) = r(s, \pi^*(s)) + \min_{p \in \mathcal{U}^{P_{s,\pi^*(s)}}(\delta)} p^\top h.$$

Fix an arbitrary $\widetilde{Q} \in \mathcal{U}^P(\delta)$. Its selected row is feasible, so

$$\min_{p \in \mathcal{U}^{P_{s,\pi^*(s)}}(\delta)} p^\top h \le \widetilde{Q}_{s,\pi^*(s)}^\top h.$$

Consequently,

$$r(s, \pi^*(s)) + \widetilde{Q}_{s,\pi^*(s)}^\top h \ge g + h(s). \tag{18}$$

Under the trajectory generated by $(\widetilde{Q}, \pi^*)$,

$$\mathbb{E}[h(S_{t+1}) \mid S_t] = \widetilde{Q}_{S_t,\pi^*(S_t)}^\top h.$$

Taking expectations in inequality (18), summing from $t = 0$ to $T-1$, and cancelling the intermediate bias terms gives

$$\mathbb{E}_{\widetilde{Q}}^{\pi^*}\left[\sum_{t=0}^{T-1} r(S_t, \pi^*(S_t)) \mid S_0 = s\right] \ge Tg + h(s) - \mathbb{E}_{\widetilde{Q}}^{\pi^*}[h(S_T) \mid S_0 = s].$$

Divide by T. Since $\mathcal{S}$ is finite, h is bounded, and therefore

$$\frac{h(s) - \mathbb{E}[h(S_T)]}{T} \longrightarrow 0.$$

It follows that

$$g_{\widetilde{Q}}^{\pi^*}(s) \geq g.$$

Because $\widetilde{Q}$ was arbitrary,

$$g_{\mathrm{rob},P}^{\pi^*}(s) = \inf_{\widetilde{Q}\in\mathcal{U}^P} g_{\widetilde{Q}}^{\pi^*}(s) \geq g.$$

For the reverse bound, choose for every state–action pair a minimizing row

$$p_{s,a}^h \in \operatorname*{arg\,min}_{p\in\mathcal{U}^{P_{s,a}}(\delta)} p^\top h.$$

These minimizers exist by compactness. Rectangularity allows us to combine them into a single kernel $Q^h \in \mathcal{U}^P(\delta)$ with $Q_{s,a}^h = p_{s,a}^h$. Since the Bellman operator takes a maximum over actions,

$$\begin{aligned} r(s,a) + (p_{s,a}^h)^\top h &\leq \max_b \{r(s,b) + (p_{s,b}^h)^\top h\} \\ &= g + h(s) \end{aligned} \tag{19}$$

for every (s,a).

Let π be any stationary randomized policy. Conditional on $S_t = s$, average inequality (19) with weights $\pi(a \mid s)$ to obtain

$$\mathbb{E}_{Q^h}^{\pi}[r(S_t,A_t) + h(S_{t+1}) \mid S_t = s] \leq g + h(s).$$

Summing and telescoping as above now gives

$$\mathbb{E}_{Q^h}^{\pi}\left[\sum_{t=0}^{T-1} r(S_t,A_t) \mid S_0 = s\right] \leq Tg + h(s) - \mathbb{E}_{Q^h}^{\pi}[h(S_T) \mid S_0 = s].$$

After division by T and passage to the upper limit,

$$g_{Q^h}^{\pi}(s) \leq g.$$

Since $Q^h \in \mathcal{U}^P$, this implies

$$g_{\mathrm{rob},P}^{\pi}(s) \leq g_{Q^h}^{\pi}(s) \leq g$$

for every stationary randomized policy π. Taking the supremum over π gives $g_{\mathrm{rob},P}^{*}(s) \leq g$. Combining this with the lower bound for π^* yields

$$g_{\mathrm{rob},P}^{\pi^*}(s) = g_{\mathrm{rob},P}^{*}(s) = g$$

for every initial state s.

It remains to prove necessity and the sharp failure assertion. Suppose $\delta \geq \delta_f^{\mathrm{bell}}(P)$. The threshold can be finite only when $f \in \mathcal{F}_{\mathrm{ne}}$. Choose disjoint nonempty I, J such that J is closable in $\mathcal{G}(P)$ and

$$\rho_{f,\mathrm{all}}^P(I) \leq \delta.$$

Let $K := \mathcal{S} \setminus (I \cup J)$. For $v \in [0,1]^K$ define

$$[Fv](s) := \max_{a\in\mathcal{A}} \min_{p\in\mathcal{U}^{P_{s,a}}(\delta)} \left\{ p(J) + \sum_{x\in K} p(x)v(x) \right\}, \qquad s \in K.$$

The row functional is continuous, so F is a continuous self-map of the compact convex cube $[0,1]^K$. Brouwer's theorem supplies a fixed point v^*. Extend it to $u \in [0,1]^{\mathcal{S}}$ by setting $u = 0$ on I and $u = 1$ on J. The δ-closed property makes $\widehat{\mathcal{L}}_P u = 0$ on I. On J, choose a nominally internal action; no creation forces every feasible row for that action to remain in J, so $\widehat{\mathcal{L}}_P u = 1$. On K the equation is exactly $Fv^* = v^*$. Hence $u = \widehat{\mathcal{L}}_P u$ is nonconstant. This proves the equivalence of (i) and (ii).

For the control obstruction, take

$$r(s,a) := \mathbf{1}\{s \in J\}.$$

Starting from I, rectangularity combines the feasible rows supported on I into a kernel that keeps the process in I under every policy, so $g_{\mathrm{rob},P}^{*}(s) = 0$ on I. Starting from J, the controller selects a nominally internal action at every state; no creation keeps every ambiguity kernel inside J, so $g_{\mathrm{rob},P}^{*}(s) = 1$ on J. Thus the robust optimal gain is state dependent. If a scalar additive eigenpair existed, the lower- and upper-drift arguments above would identify its eigenvalue with the robust optimal gain from every state, a contradiction. This proves (iii) implies (i) and completes the theorem. □

Corollary 1. For every $f \in \mathcal{F}$, $\delta_f^*(P) \le \delta_f^\dagger(P) \le \delta_f^{\mathrm{bell}}(P)$. Within the no-creation class $f \in \mathcal{F}_{\mathrm{ne}}$, either inequality can be strict; for KL both inequalities can be strict simultaneously, though strict separation need not occur for every generator in the class. In particular $\delta_f^\dagger(P) < \delta_f^{\mathrm{bell}}(P) = +\infty$ is possible. If $f \notin \mathcal{F}_{\mathrm{ne}}$, then $\delta_f^{\mathrm{bell}}(P) = +\infty$, while $\delta_f^\dagger(P)$ may be finite. When $f \in \mathcal{F}_{\mathrm{nc}} \cap \mathcal{F}_{\mathrm{ne}}$ and $\mathcal{G}(P)$ is weakly communicating, all three radii equal $+\infty$.

Proof. The first inequality is Proposition 2. If $f \notin \mathcal{F}_{\mathrm{ne}}$, then $\delta_f^{\mathrm{bell}}(P) = +\infty$ and the second inequality is immediate. Suppose $f \in \mathcal{F}_{\mathrm{ne}}$. If (I, J) is admissible in equation (4), nominal closability of J gives $\rho^P_{f,\mathrm{pol}}(J) = 0$. The same pair is therefore admissible in the definition of $\delta_f^\dagger(P)$ with objective value $\rho^P_{f,\mathrm{all}}(I)$, proving $\delta_f^\dagger(P) \le \delta_f^{\mathrm{bell}}(P)$.

For strictness take three states, one action, and, cyclically modulo three,

$$P(i \mid i) = \varepsilon, \qquad P(i+1 \mid i) = 1 - \varepsilon.$$

No proper nonempty set is closable in $\mathcal{G}(P)$, so $\delta_f^{\mathrm{bell}}(P) = +\infty$. The disjoint singletons $I = \{1\}$ and $J = \{2\}$ give $\delta_f^\dagger(P) \le \theta_f(1-\varepsilon) < +\infty$ whenever that deletion cost is finite. The creation-class conclusion follows directly from the definition of δ_f^{bell}. □

C.10 Separation from policywise unichainness

Proposition 3. *There exist P with $\mathcal{G}(P) \in \mathcal{M}$, a nonconstant reward $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, $f \in \mathcal{F}$, $\delta < \delta_f^{\mathrm{bell}}(P)$ and $h \in \mathcal{H}(P)$ such that*

$$|\operatorname{Rec}(\mathcal{G}(P)_{\pi^*})| = 2 \qquad \textit{for every } \pi^* \in \operatorname{Gr}_P(h),$$

and this occurs already when $\mathcal{U}^P(\delta) = \{P\}$ and $\delta_f^\dagger(P) = \delta_f^{\mathrm{bell}}(P) = +\infty$.

Proof. Let $\mathcal{S} = \{0,1,2\}$, $\mathcal{A} = \{\alpha, \beta\}$, and let P be the deterministic kernel

$$\begin{aligned} P(0 \mid 0,\alpha) &= 1, & P(2 \mid 0,\beta) &= 1,\\ P(1 \mid 1,\alpha) &= 1, & P(2 \mid 1,\beta) &= 1,\\ P(0 \mid 2,\alpha) &= 1, & P(1 \mid 2,\beta) &= 1, \end{aligned}$$

with $r(0,\cdot) = r(1,\cdot) = 1$ and $r(2,\cdot) = 0$, which is nonconstant.

Choose $f \in \mathcal{F}_{\mathrm{nc}} \cap \mathcal{F}_{\mathrm{ne}}$, for instance the Jeffreys generator $f(x) = (x-1)\log x$, and any finite $\delta > 0$. Each nominal row is a point mass; $f(0) = +\infty$ forbids deleting its support and $f'(\infty) = +\infty$ forbids creating any other successor, so every feasible row equals the nominal row and $\mathcal{U}^P(\delta) = \{P\}$. Since this holds at every finite radius, Theorem 2 gives $\delta_f^\dagger(P) = +\infty$.

The only nonempty closed set of $\mathcal{G}(P)$ is $\mathcal{S}$: a closed set containing 0 contains 2 through β, hence contains 1, and symmetrically from 1; a closed set containing 2 contains 0 and 1. Hence no two disjoint nonempty closed sets exist, and since a closable set that is disjoint from a closed set would contradict the same computation, $\mathcal{G}(P) \in \mathcal{M}$. Lemma 6 then gives $C(P) = \mathcal{S}$.

Take $g = 1$ and $h := (0, 0, -1)$, normalized at $s_\circ = 0$. Because $\mathcal{U}^P = \{P\}$, the robust backup is the nominal one, so

$$\begin{aligned} [\mathcal{L}_P^{\mathrm{rob}} h](0) &= \max\{1 + h(0),\, 1 + h(2)\} = 1 = g + h(0),\\ [\mathcal{L}_P^{\mathrm{rob}} h](1) &= \max\{1 + h(1),\, 1 + h(2)\} = 1 = g + h(1),\\ [\mathcal{L}_P^{\mathrm{rob}} h](2) &= \max\{0 + h(0),\, 0 + h(1)\} = 0 = g + h(2). \end{aligned}$$

Thus $(g, h) \in \operatorname{Eig}(\mathcal{L}_P^{\mathrm{rob}})$ and $h \in \mathcal{H}(P)$.

Since $h(2) = -1 < 0 = h(0) = h(1)$, the maximum at state 0 is attained only at α, and likewise at state 1. Hence every $\pi^* \in \operatorname{Gr}_P(h)$ satisfies $\pi^*(0) = \pi^*(1) = \alpha$, so $\mathcal{G}(P)_{\pi^*}(0) = \{0\}$ and $\mathcal{G}(P)_{\pi^*}(1) = \{1\}$ are two distinct minimal nonempty closed sets of $\mathcal{G}(P)_{\pi^*}$, whatever action is chosen at state 2. Therefore $|\operatorname{Rec}(\mathcal{G}(P)_{\pi^*})| = 2$ for every greedy selector. The only tie occurs at state 2, which is transient under every greedy selector, so no tie-breaking rule removes the multichain structure. □

D RECURRENT-CORE CERTIFICATION AND ONLINE REGRET

D.1 CONDITION NUMBERS AND SHARP NAVIGATION DEPENDENCE

Define the comparison class below for every kernel, and its three constants whenever that class is nonempty. For a nominally unichain policy on $R(P)$, let R_π be its recurrent class and define

$$\Pi_{\mathrm{cmp}}(P_R) := \{\pi \in \mathcal{A}^{R(P)} : \mathrm{Rec}(\mathcal{G}(P_R)_\pi) = \{R_\pi\}, \mathrm{Rec}(\mathcal{G}(\widetilde{Q})_\pi) \subseteq 2^{R_\pi} \setminus \{\varnothing\} \quad \forall \widetilde{Q} \in \mathcal{U}^{P_R}\},$$

$$D_{\mathrm{cmp}}(P_R) := \max_{\pi \in \Pi_{\mathrm{cmp}}(P_R)} \max_{\substack{x \in R(P) \\ y \in R_\pi}} \mathbb{E}^\pi_{P_R}[\tau_y \mid S_0 = x],$$

$$C_{\mathrm{occ}}(P_R) := \max_{\pi \in \Pi_{\mathrm{cmp}}(P_R),\ \widetilde{Q} \in \mathcal{U}^{P_R},\ \mu \in \Delta(R(P)):\ \mu^\top \widetilde{Q}_\pi = \mu^\top} \max_{s \in R_\pi} \frac{\mu(s)}{\mu^\pi_{P_R}(s)},$$

$$M_{\mathrm{mix}}(P_R) := \max_{\pi \in \Pi_{\mathrm{cmp}}(P_R)} \sup_{\substack{x,s \in R(P) \\ n \geq 1}} \left| \mathbb{E}^\pi_{P_R}\left[\sum_{t<n} \mathbf{1}\{S_t = s\} \mid S_0 = x\right] - n\mu^\pi_{P_R}(s) \right|,$$

where $0/0 := 0$ and $x/0 := +\infty$ for $x > 0$. The containment in the definition makes every stationary μ above supported on R_π. When $f \in \mathcal{F}_{\mathrm{ne}}$ and $\delta < \delta_f^{\mathrm{bell}}(P)$, Lemma 18 gives $\delta < \delta_f^{\mathrm{bell}}(P_R)$. Proposition 8, applied to any deterministic policy and one of its nominal recurrent classes, therefore makes $\Pi_{\mathrm{cmp}}(P_R)$ nonempty. Finiteness of this class and nominal unichainness make all three constants finite. In particular, D_{cmp} maximizes over comparison policies, unlike a diameter defined by minimizing hitting time over policies. No comparison constants or nonemptiness assumption are imposed in the creation regime.

Lemma 10.

$$C_{\mathrm{occ}}(P_R) \leq 1 + D_{\mathrm{cmp}}(P_R), \qquad M_{\mathrm{mix}}(P_R) \leq 2D_{\mathrm{cmp}}(P_R).$$

Proof. Fix $\pi \in \Pi_{\mathrm{cmp}}(P_R)$ and $s \in R_\pi$. Kac's formula and the Markov property after the first step give

$$\frac{1}{\mu^\pi_{P_R}(s)} = \mathbb{E}^\pi_s[\tau^+_s] \leq 1 + \max_x \mathbb{E}^\pi_x[\tau_s] \leq 1 + D_{\mathrm{cmp}}(P_R).$$

Every admissible adversarial stationary distribution is supported on R_π by the defining recurrent-class containment, and its mass at s is at most one. This proves the first inequality.

For the second, define $d_s(x) := \mathbf{1}\{x = s\} - \mu^\pi_{P_R}(s)$ and normalize a solution of $v_s - (P_R)_\pi v_s = d_s$ at a recurrent reference state y. The hitting-time representation gives

$$|v_s(x)| \leq \mathbb{E}^\pi_x \tau_y \leq D_{\mathrm{cmp}}(P_R).$$

Telescoping the Poisson equation therefore yields, for every x, n,

$$\left| \mathbb{E}^\pi_x\left[\sum_{t<n} d_s(S_t)\right] \right| = |v_s(x) - \mathbb{E}^\pi_x v_s(S_n)| \leq 2D_{\mathrm{cmp}}(P_R),$$

and maximizing proves the claim. □

Proposition 4. *Writing* $d_R := |R(P)|$,

$$D_{\mathrm{cmp}}(P_R) \leq (d_R - 1)p_{\min,R}(P)^{-(d_R-1)}.$$

Conversely, for every $L \geq 1$ *and* $p \in (0, 1/2]$ *there is a one-action communicating chain on* $L + 1$ *states with* $p_{\min} = p$ *and*

$$D_{\mathrm{cmp}}(P) \geq \mathbb{E}_0 \tau_L = \frac{1 - p^L}{(1-p)p^L}.$$

Hence the exponent is unavoidable for converting $(d_R, p_{\min,R})$ *to a worst-case hitting time.*

Proof. For the upper bound, every $x \in R(P)$ and $y \in R_\pi$ are joined in the unichain support graph under π by a simple path of length at most $d_R - 1$. Every path edge has probability at least $p_{\min,R}(P)$. Applying the strong Markov property in blocks of $d_R - 1$ steps gives the displayed upper bound.

For the lower bound take states $\{0, \dots, L\}$, one action, and

$$P(i+1 \mid i) = p, \qquad P(0 \mid i) = 1-p \quad (0 \le i < L), \qquad P(0 \mid L) = 1.$$

Reaching L from 0 requires L consecutive forward transitions. The first-step recursion for the waiting time of such a run gives $\mathbb{E}_0 \tau_L = (1-p^L)/((1-p)p^L)$. □

Lemma 11. *If $\delta < \delta_f^{\mathrm{bell}}(P_R)$, then $\Delta_{\mathrm{bell}}(P_R) > 0$ and*

$$\bar{\epsilon}_{\mathrm{wc}}(P_R) \ge \tfrac{1}{4}\Delta_{\mathrm{bell}}(P_R), \qquad \bar{\epsilon}_{\mathrm{face}}(P_R) \ge \tfrac{1}{16}\Delta_{\mathrm{bell}}(P_R).$$

Proof. Only $f \in \mathcal{F}_{\mathrm{ne}}$ needs proof. Throughout the ball of radius $\frac{1}{2}\min\{p_{\min,R}(P), m_{\mathrm{bell}}(P_R)\}$ around P_R on the nominal support face, every nominally positive coordinate stays at least $3p_{\min,R}(P)/4 > 0$, since an individual coordinate changes by at most half the row ℓ_1 distance. Hence the support graph, and with it $\mathfrak{W}_{\mathrm{bell}}$, agrees with that of P_R throughout this neighborhood. Since $\delta < \delta_f^{\mathrm{bell}}(P_R)$, every $(I, J) \in \mathfrak{W}_{\mathrm{bell}}(P_R)$ has a row (s, a) with

$$\theta_f(P_R(I^c \mid s, a)) > \delta, \qquad P_R(I^c \mid s, a) > q_\star.$$

The family is finite, so $m_{\mathrm{bell}}(P_R) > 0$ unless it is empty. If the obstruction family is empty, support preservation keeps it empty throughout the ball, so $\delta_f^{\mathrm{bell}}(Q) = +\infty$ there and the conclusion is immediate. Otherwise, for Q on the same support face with $\|Q - P_R\|_{\max,1} \le \frac{1}{2}\min\{p_{\min,R}(P), m_{\mathrm{bell}}(P_R)\}$, every obstruction pair retains a witness with $Q(I^c \mid s, a) > q_\star$: the maximizing row satisfies $P_R(I^c \mid s, a) \ge q_\star + m_{\mathrm{bell}}(P_R)$, and the mass of I^c changes by at most half the row ℓ_1 distance. Monotonicity of θ_f and the definition of $q_\star$ give $\rho^Q_{f,\mathrm{all}}(I) > \delta$ for every pair, hence $\delta < \delta_f^{\mathrm{bell}}(Q)$. Thus $\epsilon_{\mathrm{wc}}(P_R) \ge \frac{1}{2}\min\{p_{\min,R}(P), m_{\mathrm{bell}}(P_R)\}$. Substitution into the definitions of $\bar{\epsilon}_{\mathrm{wc}}$ and $\bar{\epsilon}_{\mathrm{face}}$, together with $\Delta_{\mathrm{bell}} \le 1 \wedge p_{\min,R} \wedge m_{\mathrm{bell}}$, gives the two displayed constants. When $f \notin \mathcal{F}_{\mathrm{ne}}$, the first bound follows from $\delta_f^{\mathrm{bell}} \equiv +\infty$ and the second from the definition of $\bar{\epsilon}_{\mathrm{face}}$. □

D.2 Parameter-bounded minimax families

Uniform exponents require bounded instance constants: even nominal average-reward learning has no uniform sublinear guarantee over kernels of unbounded diameter or bias span (Jaksch et al., 2010; Boone & Zhang, 2024). Fix S, A and $B \ge 1$. For a known reward $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$, write $H_{\mathrm{loc}}(P_R; r)$ for the local span constant generated by that reward. With $p_{\min,C}(P) := \min\{P(s' \mid s, a) > 0 : s \in C(P), a \in \mathcal{A}, s' \in \mathcal{S}\}$, define

$$\begin{aligned}
\mathfrak{P}_f(B; r) := \big\{ & P : \mathcal{G}(P) \in \mathcal{M},\ \ f \notin \mathcal{F}_{\mathrm{ne}} \Longrightarrow C(P) = \mathcal{S}, \\
& \delta < \delta_f^*(P) \le \delta_f^\dagger(P) \le \delta_f^{\mathrm{bell}}(P), \\
& \Delta_\theta(P) \ge B^{-1},\ p_{\min,C}(P) \ge B^{-1}, \\
& D_0(P) \le B,\ H_{\mathrm{loc}}(P_R; r) \le B, \\
& f \in \mathcal{F}_{\mathrm{ne}} \Longrightarrow D_{\mathrm{cmp}}(P_R) \le B \big\}, \\
\mathfrak{R}_f(T; B) := \inf_{\mathfrak{A} \in \mathfrak{A}_{\mathrm{pol}}} & \sup_{r \in [0,1]^{\mathcal{S}\times\mathcal{A}}} \sup_{s_0 \in \mathcal{S}} \sup_{P \in \mathfrak{P}_f(B; r)} \mathbb{E}^{\mathfrak{A}}_{P, r, s_0} \mathrm{Reg}_{\mathrm{pol}}(T).
\end{aligned}$$

The learner receives r before interaction and observes $s_0 = S_0$. When the reward is fixed inside an argument, we suppress it and write $\mathfrak{P}_f(B) := \mathfrak{P}_f(B; r)$ and $H_{\mathrm{loc}}(P_R) := H_{\mathrm{loc}}(P_R; r)$; the same suppression convention applies to every superscripted kernel class below. Because $D_0(P)$ maximizes over initial states in all of $\mathcal{S}$, the single condition $D_0(P) \le B$ controls entrance into $C(P)$ as well as recurrent-row coverage. No separate entrance-time parameter is used. Transient nominal states remain allowed, so this condition is strictly weaker than $C(P) = \mathcal{S}$. The communication condition is imposed only when creation is possible; the comparison-policy bound is imposed only without creation. The creation upper bound uses D_0, H_{loc}, and K_{face}, so no comparison class is needed there. For $f \in \mathcal{F}_{\mathrm{ne}}$, learning and control take place on P_R and no assumption $C(P) = \mathcal{S}$ is made. Lemma 9, including its recurrent-restriction conclusion, and inequality (21) bound all proof radii by

explicit functions of (B, f, δ). For precision, define the intermediate family and its risk by

$$\begin{aligned}\mathfrak{P}_f^\dagger(B;r) := \big\{P : \mathcal{G}(P) \in \mathcal{M},\ \ f \notin \mathcal{F}_{\mathrm{ne}} \Longrightarrow C(P) = \mathcal{S},\\ \delta_f^*(P) \le \delta < \delta_f^{\mathrm{bell}}(P),\\ \Delta_\theta(P) \ge B^{-1},\ p_{\min,C}(P) \ge B^{-1},\ \Delta_{\mathrm{bell}}(P_R) \ge B^{-1},\\ D_0(P) \le B,\ H_{\mathrm{loc}}(P_R;r) \le B,\\ f \in \mathcal{F}_{\mathrm{ne}} \Longrightarrow D_{\mathrm{cmp}}(P_R) \le B\big\},\\ \mathfrak{R}_f^\dagger(T;B) := \inf_{\mathfrak{A}\in\mathfrak{A}_{\mathrm{pol}}} \sup_{r\in[0,1]^{\mathcal{S}\times\mathcal{A}}} \sup_{s_0\in\mathcal{S}} \sup_{P\in\mathfrak{P}_f^\dagger(B;r)} \mathbb{E}^{\mathfrak{A}}_{P,r,s_0}\,\mathrm{Reg}_{\mathrm{pol}}(T).\end{aligned}$$

Thus the change from $\mathfrak{P}_f(B)$ is exactly the radius annulus and the additional explicit finite-deletion margin. Lemma 11 then derives both neighborhood proof constants from this margin. For $f \notin \mathcal{F}_{\mathrm{ne}}$ the condition adds nothing: $m_{\mathrm{bell}} = +\infty$, and the existing $p_{\min,C} \ge B^{-1}$ constraint already implies it. For $f \in \mathcal{F}_{\mathrm{ne}}$ it is exactly a margin in the finitely many deletion masses entering (4), combined with $p_{\min,R}$.

The two structurally different parts of this intermediate family are

$$\begin{aligned}\mathfrak{P}_f^{\mathrm{all}}(B;r) &:= \{P \in \mathfrak{P}_f^\dagger(B;r) : \delta < \delta_f^\dagger(P)\},\\ \mathfrak{P}_f^{\mathrm{bell}}(B;r) &:= \{P \in \mathfrak{P}_f^\dagger(B;r) : \delta \ge \delta_f^\dagger(P)\},\end{aligned}$$

with minimax risks $\mathfrak{R}_f^{\mathrm{all}}(T;B)$ and $\mathfrak{R}_f^{\mathrm{bell}}(T;B)$ defined by replacing the innermost kernel class in the preceding infimum–supremum display by $\mathfrak{P}_f^{\mathrm{all}}(B;r)$ and $\mathfrak{P}_f^{\mathrm{bell}}(B;r)$, respectively. Here all denotes the annulus in which every ambiguity kernel is weakly communicating, while bell denotes the annulus in which some ambiguity kernel is not weakly communicating but the robust Bellman equation remains solvable for every bounded reward. Keeping these families separate is essential: δ_f^* and $\delta_f^\dagger$ have different exact structural meanings even if their minimax horizon exponents ultimately coincide.

For a uniform statement, set

$$\mathfrak{P}_f^{\mathrm{core}}(B;r) := \mathfrak{P}_f(B;r), \qquad \mathfrak{R}_f^{\mathrm{core}}(T;B) := \mathfrak{R}_f(T;B),$$

$$\Xi_f := \begin{cases}\{\mathrm{core}\}, & f \in \mathcal{F}_{\mathrm{nc}},\\ \{\mathrm{core}, \mathrm{all}, \mathrm{bell}\}, & f \notin \mathcal{F}_{\mathrm{nc}},\end{cases} \qquad \alpha_f := \begin{cases}\frac{1}{2}, & f \in \mathcal{F}_{\mathrm{ne}},\\ \frac{2}{3}, & f \notin \mathcal{F}_{\mathrm{ne}}.\end{cases}$$

The restriction of Ξ_f when $f \in \mathcal{F}_{\mathrm{nc}}$ merely removes the two empty deletion annuli: under the standing family condition $\mathcal{G}(P) \in \mathcal{M}$, no deletion gives $\delta_f^*(P) = \delta_f^\dagger(P) = +\infty$ at every finite radius.

D.3 Derived proof constants and readiness certificates

For bookkeeping, the proof-level instance constant used inside several bounds remains

$$\begin{aligned}\mathfrak{b}_f(P) := D_0(P) + H_{\mathrm{loc}}(P_R) + K_{\mathrm{face}}(P_R)\\ + \begin{cases}C_{\mathrm{occ}}(P_R) + M_{\mathrm{mix}}(P_R), & f \in \mathcal{F}_{\mathrm{ne}},\\ 0, & f \notin \mathcal{F}_{\mathrm{ne}}.\end{cases}\end{aligned} \tag{20}$$

On $\mathfrak{P}_f(B)$, let $\Gamma_f(B) := \max\{\eta_f^{-1}, L_f B, B\}$. Lemmas 9 and 11 give

$$\begin{aligned}\Delta_{\mathrm{bell}}(P_R)^{-1} &\le \Gamma_f(B),\\ \bar{\epsilon}_{\mathrm{wc}}(P_R)^{-1} &\le 4\Gamma_f(B),\\ \bar{\epsilon}_{\mathrm{face}}(P_R)^{-1} &\le 16\Gamma_f(B).\end{aligned} \tag{21}$$

When creation is possible, the first inequality follows directly from $m_{\mathrm{bell}} = +\infty$ and $p_{\min,R}^{-1} \le B$; when creation is impossible, it is Lemma 9(iv). Thus the display also covers no-creation instances with transient nominal states. On the family $\mathfrak{P}_f(B)$, $D_0, H_{\mathrm{loc}} \le B$ and $p_{\min,R}^{-1} = p_{\min,C}^{-1} \le B$,

hence $K_{\mathrm{face}} \le 2B$. When $f \in \mathcal{F}_{\mathrm{ne}}$, also $D_{\mathrm{cmp}} \le B$, and Lemma 10 gives $C_{\mathrm{occ}} \le 1 + B$ and $M_{\mathrm{mix}} \le 2B$. Thus $\mathfrak{b}_f(P)$ is bounded by an explicit function of B in both regimes.

For the no-creation algorithm, B_k is the strongly connected component of the observed state graph containing S_{t_k}, $q^{\uparrow}$ denotes zero extension from B_k to $\mathcal{S}$, and

$$\mathcal{C}^{\circ}_{k,B_k}(s,a) := \{q \in \Delta(B_k) : q^{\uparrow} \in \mathcal{C}_k(s,a),\ \mathrm{supp}(q) \subseteq \widehat{G}_k(s,a) \cap B_k\}.$$

The history-measurable readiness condition is defined using

$$\mathsf{CoreReady}_k(B) :\Longleftrightarrow B \in \mathfrak{B}_k \quad \text{or} \quad \big(q_\star = 0,\ B = \mathcal{S},\ \widehat{G}_k|_B \in \mathcal{M}(B)\big).$$

The second branch is a one-sided certificate, indeed observed edges are genuine and all genuine edges belong to the robust core when $q_\star = 0$. On this branch the observed graph $\widehat{G}_k|_{B_k}$ replaces $H_k^-|_{B_k}$ as the certified mandatory graph in the completion below, and the optimistic centers are restricted to the observed support, $\mathrm{supp}(Q_{s,a}) = \widehat{G}_k(s,a) \cap B_k$ for every $(s,a) \in B_k \times \mathcal{A}$. Since $q_\star = 0$, every observed edge is then a core edge of every admissible center, so the hypotheses of Proposition 7 hold even when H_k^- has empty rows. Observing an edge once certifies its existence, not that the lower bound over the still wide confidence ball exceeds $q_\star$, so the graph replacement alone would not suffice. The full readiness condition is

$$\begin{aligned}\mathsf{CoreReady}_k(B_k), \quad H_k^-|_{B_k} \subseteq \widehat{G}_k|_{B_k}, \qquad \prod_{s\in B_k, a\in\mathcal{A}} \mathcal{C}^{\circ}_{k,B_k}(s,a) \neq \varnothing,\\ \delta < \inf_{Q \in \prod_{s,a} \mathcal{C}^{\circ}_{k,B_k}(s,a)} \delta_f^{\mathrm{bell}}(Q).\end{aligned} \tag{22}$$

For the stronger completion used throughout the no-creation Bellman regime, define, for every candidate recurrent set B, the history-measurable certificate

$$\begin{aligned}\mathsf{BellReady}_k(B) :\Longleftrightarrow \widehat{G}_k|_B \in \mathcal{M}(B), \qquad \widehat{G}_k(s,a) \subseteq B \quad \forall (s,a) \in B \times \mathcal{A},\\ \prod_{s\in B, a\in\mathcal{A}} \mathcal{C}^{\circ}_{k,B}(s,a) \neq \varnothing,\\ \mathrm{supp}(q) = \widehat{G}_k(s,a) \quad \forall (s,a) \in B \times \mathcal{A},\ \forall q \in \mathcal{C}^{\circ}_{k,B}(s,a),\\ \text{and, for every } I, J \subseteq B \text{ nonempty with } I \cap J = \varnothing \text{ and } J \text{ closable in } \widehat{G}_k|_B,\\ \exists (s,a) \in I \times \mathcal{A} : \quad \inf_{q \in \mathcal{C}^{\circ}_{k,B}(s,a)} q(B \setminus I) > q_\star.\end{aligned} \tag{23}$$

The disjoint-pair quantifiers in (23) range over finitely many discrete choices, while the row quantifiers range over the convex confidence sets $\mathcal{C}^{\circ}_{k,B}(s,a)$. Each instance of the last line is a finite-dimensional optimization of the linear functional $q \mapsto q(B \setminus I)$ over a convex set, so the certificate is evaluated by finitely many finite-dimensional convex-optimization checks. our rate claim concerns measurability and does not claim that this implementation is polynomial time. Define the support-recovery threshold

$$n_{\mathrm{supp}}(P,\alpha) := \left\lceil \frac{1}{p_{\min,R}(P)} \log \frac{6 d_R^2 A}{\alpha} \right\rceil,$$

and the Bellman-completion threshold

$$n_{\mathrm{BC}}(P_R,\alpha) := n_{\mathrm{supp}}(P,\alpha) \vee \min\left\{n \ge 1 : 2\beta_{\bar\alpha}(n) \le \tfrac{1}{16}\Delta_{\mathrm{bell}}(P_R)\right\}.$$

On the joint confidence and exploration event, if $f \in \mathcal{F}_{\mathrm{ne}}$ and $n_T^{\mathrm{OC}}(P,\alpha) \ge n_{\mathrm{BC}}(P_R,\alpha)$, the first post-exploration episode has $B_k = C(P)$, support recovery gives $\widehat{G}_k|_{B_k} = \mathcal{G}(P_R)$, and $\mathsf{BellReady}_k(B_k)$ holds there and thereafter. Indeed, every obstruction pair for P_R has a witness whose true outgoing mass exceeds $q_\star$ by at least $m_{\mathrm{bell}}(P_R)$, while the total-variation error of any confidence row is at most one half of its ℓ_1 error. With $m_T^{\mathrm{OC}} := \lceil \sqrt{T} \rceil$, set

$$\begin{aligned}n_T^{\mathrm{OC}}(P,\alpha) &:= \max\{n \in \mathbb{N}_0 : \mathfrak{t}_P(n,\alpha) \le m_T^{\mathrm{OC}}\},\\ n^{\sharp}_{\mathrm{cert}}(P,\alpha) &:= \begin{cases} n_{\mathrm{supp}}(P,\alpha), & q_\star = 0,\ C(P) = \mathcal{S},\\ \min\{n \ge 1 : 2\beta_{\bar\alpha}(n) < \epsilon_{\mathrm{top}}(P)\}, & \text{otherwise,}\end{cases}\\ n_{\mathrm{ne}}(P,\alpha) &:= n^{\sharp}_{\mathrm{cert}}(P,\alpha) \vee \min\{n \ge 1 : 2\beta_{\bar\alpha}(n) \le \bar\epsilon_{\mathrm{face}}(P_R)\} \vee n_{\mathrm{supp}}(P,\alpha),\\ \Upsilon_{\mathrm{ne}}(P) &:= \begin{cases} p_{\min,C}(P)^{-1}, & q_\star = 0,\ C(P) = \mathcal{S},\\ \epsilon_{\mathrm{top}}(P)^{-2}, & \text{otherwise}\end{cases} + \bar\epsilon_{\mathrm{face}}(P_R)^{-2} + p_{\min,C}(P)^{-1}.\end{aligned}$$

Then $\Upsilon_{\mathrm{ne}}(P) = O_{f,\delta}(B^2)$ uniformly on the subfamily where $C(P) = \mathcal{S}$ or $q_\star > 0$. The remaining case $q_\star = 0$, $C(P) \neq \mathcal{S}$ is handled by the Bellman certificate and $n_{\mathrm{BC}}(P_R, \alpha)$ above. On the joint confidence and exploration event, $n_T^{\mathrm{OC}} \geq n_{\mathrm{ne}}$ implies that the first post-exploration episode has $B_k = C(P)$ and satisfies condition (22).

D.4 Anytime row confidence

Lemma 12. $\mathbb{P}_P(\mathcal{E}_{\mathrm{conf}}) \geq 1 - \bar{\alpha}$.

Proof. Fix a state–action pair (s, a) and let

$$\tau_1 < \tau_2 < \cdots$$

be the successive times at which $(S_t, A_t) = (s, a)$. Define $Y_m := S_{\tau_m+1}$. The definition of the controlled process gives, conditional on the history through time τ_m,

$$\mathbb{P}(Y_m = s' \mid \mathcal{F}_{\tau_m}) = P(s' \mid s, a). \tag{24}$$

The right side does not depend on the past. Iterating conditional expectations in equation (24) shows that $(Y_m)_{m\geq 1}$ are independent and identically distributed with common law $P_{s,a}$. Thus the empirical row after the nth visit is an ordinary multinomial empirical distribution, even though the visit times themselves are adaptive.

For a multinomial distribution on S outcomes, a Weissman bound gives

$$\mathbb{P}\left(\|\widehat{P}_{n,s,a} - P_{s,a}\|_1 > \varepsilon\right) \leq (2^S - 2)e^{-n\varepsilon^2/2}, \qquad \varepsilon > 0. \tag{25}$$

If $\beta_{\bar{\alpha}}(n) = 2$, the event with strict inequality $\|\widehat{P} - P\|_1 > 2$ is empty because the ℓ_1 distance between two probability vectors is at most two. Otherwise,

$$\beta_{\bar{\alpha}}(n) = \sqrt{\frac{2\log(c_{\bar{\alpha}} n(n+1))}{n}}.$$

Substitution into inequality (25) gives

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{P}_{n,s,a} - P_{s,a}\|_1 > \beta_{\bar{\alpha}}(n)\right) \\
&\quad \leq (2^S - 2)\exp\{-\log(c_{\bar{\alpha}} n(n+1))\} \\
&\quad = \frac{2^S - 2}{c_{\bar{\alpha}} n(n+1)} \\
&\quad = \frac{\bar{\alpha}}{SAn(n+1)}.
\end{aligned}$$

Take a union bound over all $n \geq 1$, all $s \in \mathcal{S}$, and all $a \in \mathcal{A}$. Since

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1}\right) = 1,$$

the probability that any one of these confidence balls ever excludes its true row is at most

$$\sum_{s,a} \sum_{n=1}^{\infty} \frac{\bar{\alpha}}{SAn(n+1)} = \bar{\alpha}.$$

When a row has not yet been sampled, its radius is 2 and its ball contains the entire simplex, so it also contains the true row. Therefore, on the complementary event of probability at least $1 - \bar{\alpha}$, every ball $\mathcal{B}_j(s, a)$ contains $P_{s,a}$. An intersection of sets that all contain $P_{s,a}$ also contains it, so

$$P_{s,a} \in \mathcal{C}_k(s, a)$$

simultaneously for all episodes and all rows. □

D.5 Recurrent-core certification

Proposition 5. *On $\mathcal{E}_{\mathrm{conf}}$,*

$$H_k^- \subseteq \mathcal{G}_\wedge(P) \subseteq H_k^+.$$

Every certified recurrent region B_k is closed in $\mathcal{G}_\wedge(P)$. If in addition $\delta < \delta_f^*(P)$, so that $C_\wedge(P)$ is defined, then $C_\wedge(P) \subseteq B_k$. Thus certification requires no transition samples from $\mathcal{S} \setminus B_k$.

Proof. Work on the event from Lemma 12. Fix an episode k, a row (s, a), and a successor s'. If $s' \in H_k^-(s, a)$, then by definition

$$\inf_{q \in \mathcal{C}_k(s,a)} \theta_f(q(s')) > \delta.$$

The true row belongs to the confidence set, so

$$\theta_f(P(s' \mid s, a)) \geq \inf_{q \in \mathcal{C}_k(s,a)} \theta_f(q(s')) > \delta.$$

The robust core characterization in Theorem 1 therefore gives

$$s' \in \mathcal{G}_\wedge(P)(s, a).$$

This proves

$$H_k^-(s, a) \subseteq \mathcal{G}_\wedge(P)(s, a).$$

Conversely, if $s' \in \mathcal{G}_\wedge(P)(s, a)$, then

$$\theta_f(P(s' \mid s, a)) > \delta.$$

Since $P_{s,a} \in \mathcal{C}_k(s, a)$,

$$\sup_{q \in \mathcal{C}_k(s,a)} \theta_f(q(s')) \geq \theta_f(P(s' \mid s, a)) > \delta,$$

and hence $s' \in H_k^+(s, a)$. Thus

$$H_k^-(s, a) \subseteq \mathcal{G}_\wedge(P)(s, a) \subseteq H_k^+(s, a)$$

for every row, proving the graph inclusions.

Now let B_k be certified. For every $s \in B_k$ and every action a, the certificate and the upper inclusion give

$$\mathcal{G}_\wedge(P)(s, a) \subseteq H_k^+(s, a) \subseteq B_k.$$

This is exactly the definition that B_k is closed in $\mathcal{G}_\wedge(P)$.

Apply Lemma 6(ii) to $G = \mathcal{G}_\wedge(P)$ and $B = B_k$. Under $\delta < \delta_f^*(P)$, the core graph is weakly communicating and its recurrent set is $C_\wedge(P)$. Hence

$$C_\wedge(P) \subseteq B_k.$$

The proof used only confidence sets for rows based at states in B_k; rows based outside B_k play no role in the certificate or in the proof. □

D.6 Completion from a certified recurrent region

Lemma 13. *If a finite directed graph F has a unique sink strongly connected component and $F \subseteq F'$, then F' also has a unique sink strongly connected component.*

Proof. Let K be any sink component of F' and choose $x \in K$. In F, every vertex has a directed path to the unique sink component R_F. The same path lies in F'. Since K is closed in F' and the path starts at $x \in K$, it ends in K, so $K \cap R_F \neq \varnothing$. Any two sink components of F' would therefore both meet R_F. But R_F remains strongly connected in F', so all its vertices belong to a single strongly connected component of F'. The two sink components must therefore coincide. A finite directed graph has at least one sink strongly connected component, which proves existence and uniqueness. □

Lemma 14. *Let $G \in \mathcal{M}$, let $H^- \subseteq G \subseteq H^+$, and suppose a nonempty B satisfies*

$$H^+(s,a) \subseteq B \quad (s \in B, a \in \mathcal{A}), \qquad H^-|_B \in \mathcal{M}(B).$$

For every $\pi \in \Pi_{\mathrm{d}}$ and every $R \in \mathrm{Rec}((H^-|_B)_\pi)$, there is a $\bar{\pi} \in \Pi_{\mathrm{d}}$ that agrees with π on R and is unichain in G. Moreover, $\bar{\pi}$ is unichain in every transition graph containing G.

Proof. Write $H := H^-|_B$.

We first prove that B contains the recurrent set of G. For every $s \in B$ and $a \in \mathcal{A}$,

$$G(s,a) \subseteq H^+(s,a) \subseteq B.$$

Thus B is closed in G. Lemma 6(ii) therefore gives

$$C(G) \subseteq B.$$

We next construct a policy that reaches R inside H. Choose any Markov kernel K whose support graph is H and whose probabilities are strictly positive on every edge of H; for instance, make each row uniform on its support. Since R is a recurrent class of H_π, it is strongly connected and

$$H(s,\pi(s)) \subseteq R, \qquad s \in R.$$

Thus R is closable in H.

For $x \in B$, let

$$v(x) := \sup_\sigma \mathbb{P}^\sigma_K(\tau_R < \infty \mid S_0 = x),$$

where the supremum ranges over all policies and $\tau_R := \inf\{t \geq 0 : S_t \in R\}$. The reachability dynamic-programming equations are

$$\begin{aligned} v(x) &= 1, && x \in R, \\ v(x) &= \max_{a \in \mathcal{A}} \sum_y K(y \mid x,a) v(y), && x \in B \setminus R. \end{aligned}$$

We show that $v(x) = 1$ for every $x \in B$. Suppose otherwise and set

$$m := \min_{x \in B} v(x) < 1, \qquad I := \{x \in B : v(x) = m\}.$$

The set I is nonempty and disjoint from R. Fix $x \in I$. For every action a, the Bellman equation and the fact that $v(x)$ is the maximum over actions give

$$\sum_y K(y \mid x,a) v(y) \leq v(x) = m.$$

But $v(y) \geq m$ for every y, so the left side is a convex combination of numbers all at least m. It can be at most m only if

$$K(y \mid x,a) > 0 \quad \Longrightarrow \quad v(y) = m.$$

Since K is positive on every edge of H, this implication is equivalent to

$$H(x,a) \subseteq I \qquad (x \in I,\ a \in \mathcal{A}).$$

Thus I is closed in H. We already observed that R is closable in H. The disjoint nonempty sets I and R contradict the weak communication of H. Therefore $v \equiv 1$.

We also make explicit why one deterministic stationary policy attains this almost-sure reachability value. For $\gamma \in (0,1)$, define

$$v_\gamma(x) := \sup_\sigma \mathbb{E}^\sigma_K[\gamma^{\tau_R} \mathbf{1}\{\tau_R < \infty\} \mid S_0 = x].$$

The discounted reachability operator is a contraction on $B \setminus R$, so an action attaining its finite maximum at each state defines a deterministic stationary optimizer σ_γ. As $\gamma \uparrow 1$, monotone convergence gives $v_\gamma(x) \uparrow v(x) = 1$. There are only finitely many deterministic stationary selectors.

Hence along some sequence $\gamma_n \uparrow 1$, the same selector, call it σ, equals σ_{γ_n}. For this fixed selector and every x,

$$\begin{aligned} 1 &= \lim_n v_{\gamma_n}(x) \\ &= \lim_n \mathbb{E}_K^{\sigma}[\gamma_n^{\tau_R} \mathbf{1}\{\tau_R < \infty\} \mid S_0 = x] \\ &= \mathbb{P}_K^{\sigma}(\tau_R < \infty \mid S_0 = x), \end{aligned}$$

where the last equality is monotone convergence. Modify σ on R by setting $\sigma(s) = \pi(s)$ there. This does not change the hitting event, and it makes R closed and irreducible under the resulting selector. Therefore R is the unique recurrent class of H_σ: a second recurrent class outside R would prevent its states from reaching R with probability one.

We now pass from H to G. Extend σ arbitrarily to $\mathcal{S} \setminus B$ and call the resulting policy $\bar{\pi}$. By construction, $\bar{\pi} = \pi$ on R.

Apply Lemma 13 first on state space B with $F = H_\sigma$ and $F' = G_{\bar{\pi}}|_B$. Since $H \subseteq G$, the latter is a supergraph and has one recurrent class. Because B is closed in G, that class is also closed in the full policy graph $G_{\bar{\pi}}$. No second recurrent class can lie outside B: every state recurrent under a deterministic policy belongs to $C(G)$, and $C(G) \subseteq B$. Hence $G_{\bar{\pi}}$ has exactly one recurrent class, so $\bar{\pi}$ is unichain in G.

Finally, we prove stability under arbitrary supergraphs. Let G' be any transition graph with $G \subseteq G'$. Then $G_{\bar{\pi}} \subseteq G'_{\bar{\pi}}$. We just proved that $G_{\bar{\pi}}$ has a unique sink strongly connected component. Lemma 13 shows that $G'_{\bar{\pi}}$ also has a unique sink component. Thus the same completed policy is unichain in every transition graph containing G. □

D.7 Finite recurrent-core sample complexity

Proposition 6. *Suppose $f \in \mathcal{F}$ and $\delta < \delta_f^*(P)$. Then $\epsilon_e(P) > 0$ for every robust core edge e based in $C(P)$. If*

$$\epsilon_{s,a}(P) > 0, \qquad s \in C(P),\ a \in \mathcal{A},$$

then every $n_{s,a}^{\mathrm{top}}(P, \alpha)$ is finite. On $\mathcal{E}_{\mathrm{conf}}$, if

$$N_k(s,a) \geq n_{s,a}^{\mathrm{top}}(P, \alpha), \qquad s \in C(P),\ a \in \mathcal{A},$$

then $B = C(P)$ satisfies condition (8)*, and consequently $\tau_{\mathrm{topo}} \leq t_k$.*

Proof. We first prove positivity of the radii. Let $e = (s, a, s')$ be a robust core edge based in $C(P)$. The core characterization gives the strict inequality

$$\theta_f(P(e)) > \delta.$$

Choose a number c_e strictly between them:

$$\delta < c_e < \theta_f(P(e)).$$

Lower semicontinuity of θ_f at $P(e)$ implies that there exists $\eta_e > 0$ such that

$$|x - P(e)| < \eta_e \quad \Longrightarrow \quad \theta_f(x) > c_e > \delta.$$

Thus every sufficiently small positive $\epsilon < \eta_e$ belongs to the set defining $\epsilon_e(P)$, and hence

$$\epsilon_e(P) > 0.$$

Now assume the additional condition in Proposition 6,

$$\epsilon_{s,a}(P) > 0 \qquad (s \in C(P),\ a \in \mathcal{A}).$$

Since $\beta_{\bar{\alpha}}(n) \to 0$, a finite integer satisfying $2\beta_{\bar{\alpha}}(n) < \epsilon_{s,a}(P)$ exists. The confidence radius is nonincreasing in n: after the possible initial truncation at two, the derivative of $n \mapsto \log(c_{\bar{\alpha}} n(n+1))/n$ is negative because $\log(c_{\bar{\alpha}} n(n+1)) > 2$ whereas $(2n+1)/(n+1) < 2$. Hence

$$N_k(s,a) \geq n_{s,a}^{\mathrm{top}} \quad \Longrightarrow \quad 2\beta_{\bar{\alpha}}(N_k(s,a)) < \epsilon_{s,a}(P).$$

Now work on the anytime confidence event and fix a row (s, a) with $s \in C(P)$. Every $q \in \mathcal{C}_k(s, a)$ belongs, in particular, to the current ball $\mathcal{B}_k(s, a)$. Both q and $P_{s,a}$ are within $\beta_{\bar{\alpha}}(N_k(s, a))$ of the empirical row. Therefore

$$\begin{aligned}\|q - P_{s,a}\|_1 &\le \|q - \widehat{P}_{k,s,a}\|_1 + \|\widehat{P}_{k,s,a} - P_{s,a}\|_1 \\ &\le 2\beta_{\bar{\alpha}}(N_k(s, a)) \\ &< \epsilon_{s,a}(P). \end{aligned} \tag{26}$$

Let $s' \in \mathcal{G}_\wedge(P)(s, a)$ and write $e = (s, a, s')$. Since $\epsilon_{s,a}(P) \le \epsilon_e(P)$, inequality (26) gives

$$|q(s') - P(e)| < \epsilon_e(P). \tag{27}$$

By Lemma 9(i) the set of radii in the definition of $\epsilon_e(P)$ is the interval $(0, P(e) - q_\star)$, up to its right endpoint; in particular it is downward closed, so every radius strictly below $\epsilon_e(P)$ has the defining property, and inequality (27) implies

$$\theta_f(q(s')) > \delta \qquad \text{for every } q \in \mathcal{C}_k(s, a).$$

Taking the infimum over the confidence set gives

$$\inf_{q \in \mathcal{C}_k(s,a)} \theta_f(q(s')) > \delta,$$

so $s' \in H_k^-(s, a)$. Thus

$$\mathcal{G}_\wedge(P)(s, a) \subseteq H_k^-(s, a) \qquad (s \in C(P)).$$

Proposition 5 already gives the reverse inclusion $H_k^- \subseteq \mathcal{G}_\wedge(P)$. Therefore

$$H_k^-|_{C(P)} = \mathcal{G}_\wedge(P)|_{C(P)}.$$

Next let $s' \notin C(P)$. The recurrent set $C(P)$ is closed under every action in the nominal graph, so

$$P(s' \mid s, a) = 0.$$

Using inequality (26) and $\epsilon_{s,a}(P) \le \epsilon_0(P)$,

$$0 \le q(s') < \epsilon_0(P).$$

By Lemma 9(i) the set of radii in the definition of $\epsilon_0(P)$ is likewise downward closed, with supremum $q_\star$. Hence

$$\theta_f(q(s')) \le \delta \qquad \text{for every } q \in \mathcal{C}_k(s, a).$$

It follows that

$$\sup_{q \in \mathcal{C}_k(s,a)} \theta_f(q(s')) \le \delta,$$

so $s' \notin H_k^+(s, a)$. We have proved

$$H_k^+(s, a) \subseteq C(P) \qquad (s \in C(P),\ a \in \mathcal{A}).$$

Finally, $C(P)$ is a closed set containing $C_\wedge(P)$. Restricting the weakly communicating graph $\mathcal{G}_\wedge(P)$ to such a set preserves its recurrent set $C_\wedge(P)$; every state in $C(P) \setminus C_\wedge(P)$ remains transient under every policy. Hence

$$\mathcal{G}_\wedge(P)|_{C(P)} \in \mathcal{M}(C(P)).$$

Together with the two identities above, this gives

$$\begin{aligned} &H_k^+(s, a) \subseteq C(P), && s \in C(P),\ a \in \mathcal{A}, \\ &H_k^-|_{C(P)} \in \mathcal{M}(C(P)). \end{aligned}$$

These are exactly the two certificate conditions with $B = C(P)$. Therefore the certificate holds at episode k, and by the definition of τ_{topo} we have $\tau_{\text{topo}} \le t_k$. □

D.8 HIGH-PROBABILITY TOPOLOGY TIME

Theorem 6. *Suppose $f \in \mathcal{F}$, $\delta < \delta_f^*(P)$, and $M_{\mathrm{top}}(P,\alpha) < +\infty$. If the learner executes π_0 until condition* (8) *first holds, checking the certificate after every transition, then with probability at least $1-\alpha$,*

$$\tau_{\mathrm{topo}} \le 2D_0(P)M_{\mathrm{top}}(P,\alpha)\left\lceil \log_2 \frac{3M_{\mathrm{top}}(P,\alpha)}{\alpha} \right\rceil,$$

a bound that is finite and independent of the horizon T. In the notation above this reads

$$\tau_{\mathrm{topo}} = \widetilde{O}\left(\frac{D_0(P)S|C(P)|A}{\epsilon_{\mathrm{top}}(P)^2}\right),$$

and, with the divergence-only constants η_f and L_f of Lemma 9(ii),

$$\tau_{\mathrm{topo}} = \widetilde{O}\left(D_0(P)S|C(P)|A \max\left\{\frac{1}{\epsilon_0(P)^2}, \frac{1}{\eta_f^2}, \frac{L_f^2}{\Delta_\theta(P)^2}, \frac{1}{p_{\min,C}(P)^2}\right\}\right),$$

where $\widetilde{O}$ suppresses logarithms in $S, A, 1/\alpha$ and $1/\epsilon_{\mathrm{top}}(P)$; the linear factor S is kept explicit because $c_{\bar\alpha}$ is exponential in S. We use $1/(+\infty)^2 := 0$, the term $\epsilon_0(P)^{-2}$ is absent when $C(P) = \mathcal{S}$ and equals $q_\star^{-2}$ otherwise. Only the last two entries depend on P, and both are constrained directly by the minimax families of Section 6.

Proof. Set

$$M := M_{\mathrm{top}}(P,\alpha), \qquad m_0 := \left\lceil \log_2 \frac{3M}{\alpha} \right\rceil.$$

The assumption $M < +\infty$ implies that every summand $n_{s,a}^{\mathrm{top}}(P,\alpha)$ is finite. By its definition, for each recurrent row there is therefore a finite n satisfying

$$0 < 2\beta_{\bar\alpha}(n) < \epsilon_{s,a}(P),$$

so $\epsilon_{s,a}(P) > 0$. Hence the nonvacuous part of Proposition 6 is available once the required sample counts have been accumulated. The anytime confidence event has probability at least $1-\bar\alpha = 1-\alpha/3$.

We first record the stopping-time block estimate used throughout the sequel. Suppose τ is a hitting time for a Markov chain and, from every state in the relevant region,

$$\mathbb{E}_x[\tau] \le D < +\infty.$$

If $D = 0$, then $\tau = 0$ almost surely. Otherwise, in the applications below τ is integer valued and the presence of a state outside the target implies $D \ge 1$. Set the integer block length

$$w := \lfloor 2D \rfloor.$$

Since $\{\tau > w\} = \{\tau \ge w+1\}$, Markov's inequality in its integer-valued form gives $\mathbb{E}_x[\tau] \ge (w+1)\mathbb{P}_x(\tau \ge w+1)$ and therefore

$$\mathbb{P}_x(\tau > w) \le \frac{\mathbb{E}_x[\tau]}{w+1} \le \frac{D}{w+1} < \frac{1}{2},$$

because $w + 1 > 2D$. Conditional on the history at the end of a failed block, the remaining hitting problem starts from the current state and again has conditional mean at most D. The preceding one-block bound is therefore at most $1/2$ after every failed block. Induction over m blocks gives

$$\sup_x \mathbb{P}_x(\tau > wm) \le 2^{-m}.$$

Since $w \le 2D$, the event $\{\tau > 2Dm\}$ is contained in $\{\tau > wm\}$, and hence

$$\sup_x \mathbb{P}_x(\tau > 2Dm) \le 2^{-m}, \qquad m \ge 1. \tag{28}$$

Construct a deterministic list of M requested samples by listing each pair (s,a) with $s \in C(P)$ exactly $n_{s,a}^{\mathrm{top}}(P,\alpha)$ times. Let the list be

$$(s_1,a_1),\dots,(s_M,a_M).$$

Starting from the arbitrary initial state, wait for the first listed pair, take its transition, and then continue with the second listed pair. A request is complete only after the successor of its target action has been observed. Denote the successive waiting times to completion by $W_1, \dots, W_M$.

From every state in $\mathcal{S}$, the expected time to hit any requested pair is at most $D_0(P) - 1$ by the global definition of $D_0(P)$, observing its successor takes one further transition. The strong Markov property therefore gives, conditionally on the entire past before each waiting period,

$$\mathbb{E}[W_\ell \mid \text{past}] \le D_0(P), \qquad \ell = 1, \dots, M.$$

Inequality (28), conditional on the history at the beginning of each waiting period, gives

$$\mathbb{P}\left(W_\ell > 2D_0(P)m_0 \mid \text{past before } W_\ell\right) \le 2^{-m_0} \le \frac{\alpha}{3M}.$$

Taking expectations removes the conditioning, and a union bound over the M requested visits yields

$$\mathbb{P}\left(\exists\, \ell :\ W_\ell > 2D_0(P)m_0\right) \le M\frac{\alpha}{3M} = \frac{\alpha}{3}.$$

On the complementary event, all requested visits occur within

$$\sum_{\ell=1}^{M} W_\ell \le 2D_0(P)Mm_0$$

steps from the initial state. Thus every recurrent pair has accumulated at least its required number of samples.

Intersect the anytime confidence event and the event of timely completion of all requested visits. Their combined failure probability is at most $2\alpha/3 \le \alpha$. On this intersection, Proposition 6 certifies $B = C(P)$ as soon as the last requested visit occurs. Because the algorithm checks the certificate after every transition,

$$\tau_{\text{topo}} \le 2D_0(P)M_{\text{top}}(P,\alpha)\left\lceil \log_2 \frac{3M_{\text{top}}(P,\alpha)}{\alpha} \right\rceil. \tag{29}$$

Every quantity on the right depends only on the instance and confidence level, not on the horizon T.

It remains to derive the two displayed asymptotic forms. Fix a recurrent row (s,a) and abbreviate

$$\epsilon := \epsilon_{s,a}(P), \qquad c := c_{\bar{\alpha}}.$$

The defining inequality for the sampling threshold is

$$2\beta_{\bar{\alpha}}(n) < \epsilon.$$

When the confidence radius is not truncated at two, squaring both sides shows that it is enough to have

$$\frac{8\log(cn(n+1))}{n} < \epsilon^2. \tag{30}$$

To make the order explicit, set

$$\bar{\epsilon} := \min\{1, \epsilon\}, \qquad \Lambda := \log\left(\frac{64c}{\bar{\epsilon}^4}\right),$$

and choose

$$n := \left\lceil \frac{48\Lambda}{\bar{\epsilon}^2} \right\rceil.$$

Since $\Lambda \ge \log 64 > 4$ and $\bar{\epsilon} \le 1$, we have $\Lambda/\bar{\epsilon}^2 \ge 1$ and therefore

$$n \le \frac{49\Lambda}{\bar{\epsilon}^2}, \qquad n+1 \le 2n.$$

Consequently,

$$\begin{aligned}\log(cn(n+1)) &\le \log(2c) + 2\log n \\ &\le \log(2c) + 2\log 49 + 2\log\Lambda + 4\log(1/\bar{\epsilon}).\end{aligned}$$

Each group is controlled by Λ: from the definition of Λ,

$$\log(2c) \le \Lambda,$$
$$2\log 49 \le 2\Lambda,$$
$$2\log\Lambda \le \Lambda,$$
$$4\log(1/\bar{\epsilon}) \le \Lambda.$$

Thus

$$\log(cn(n+1)) \le 5\Lambda.$$

It follows that

$$\frac{8\log(cn(n+1))}{n} \le \frac{40\Lambda}{48\Lambda/\bar{\epsilon}^2} = \frac{5}{6}\bar{\epsilon}^2 \le \epsilon^2.$$

The factor $5/6$ makes the inequality strict. Therefore

$$\begin{aligned} n_{s,a}^{\mathrm{top}}(P,\alpha) &= O\!\left(\frac{1}{\min\{1,\epsilon_{s,a}(P)\}^2}\log\frac{c_{\bar{\alpha}}}{\min\{1,\epsilon_{s,a}(P)\}^4}\right) \\ &= \widetilde{O}\big(\epsilon_{s,a}(P)^{-2}\big), \end{aligned}$$

where the second line suppresses logarithms and treats radii larger than one as constants.

By definition,

$$\epsilon_{s,a}(P) \ge \epsilon_{\mathrm{top}}(P) \qquad (s \in C(P),\ a \in \mathcal{A}).$$

There are $|C(P)|A$ recurrent state–action pairs, so summing their thresholds gives

$$M_{\mathrm{top}}(P,\alpha) = \sum_{s\in C(P),a\in\mathcal{A}} n_{s,a}^{\mathrm{top}}(P,\alpha) \tag{31}$$

$$= \widetilde{O}\!\left(\frac{S|C(P)|A}{\epsilon_{\mathrm{top}}(P)^2}\right). \tag{32}$$

Substituting equation (31) into inequality (29), and absorbing the two logarithmic block factors into $\widetilde{O}$, yields

$$\tau_{\mathrm{topo}} = \widetilde{O}\!\left(\frac{D_0(P)S|C(P)|A}{\epsilon_{\mathrm{top}}(P)^2}\right).$$

It remains to convert this into the margin form. By Lemma 9(i) every core edge has $\epsilon_e(P) = P(e) - q_\star$, and by Lemma 9(ii) there are $\eta_f \in (0,1)$ and $L_f \in [1,\infty)$, depending only on (f,δ), with

$$\epsilon_{\mathrm{top}}(P) \ge \min\left\{\epsilon_0(P),\ \eta_f,\ \frac{\Delta_\theta(P)}{L_f},\ p_{\min,C}(P)\right\},$$

where the term $\epsilon_0(P)$ is present only when $C(P) \ne \mathcal{S}$ and equals $q_\star$ there. In Theorem 6 we write $\rho := \eta_f$ and $L := L_f$ for the local modulus, and we note that no separate treatment of $f(0) = +\infty$ is needed: in that case $q_\star = 0$, so $\epsilon_e(P) = P(e)$ and the third term $p_{\min,C}(P)$ is the operative one, while $\Delta_\theta(P) = +\infty$. For positive extended-real x, y,

$$\frac{1}{\min\{x,y\}^2} = \max\left\{\frac{1}{x^2}, \frac{1}{y^2}\right\},$$

with $1/(+\infty)^2 = 0$. Therefore

$$\frac{1}{\epsilon_{\mathrm{top}}(P)^2} \le \max\left\{\frac{1}{\epsilon_0(P)^2}, \frac{1}{\rho^2}, \frac{L^2}{\Delta_\theta(P)^2}, \frac{1}{p_{\min,C}(P)^2}\right\}.$$

Substitution into the first burn-in bound proves the second display. □

A uniformly valid certificate is *one-edge-decisive at mass p* if two admissible instances have identical observation laws except that the indicator of one edge is Bernoulli with mean 0 versus p, and validity requires opposite certificate decisions.

Theorem 7. *Suppose $q_\star(f,\delta)=0$, and let $G_0 \subseteq \mathcal{G}(P)$ be a weakly communicating witness subgraph with m edges, each having probability at least $p_{\min,G_0} > 0$. Once every row containing a witness edge has at least*

$$n_{\mathrm{disc}} = \left\lceil p_{\min,G_0}^{-1} \log(m/\alpha) \right\rceil$$

samples, the observed-edge graph contains G_0 with probability at least $1-\alpha$. This order is sharp for certificates one-edge-decisive at mass $p_{\min,G_0}$: any uniformly valid such certificate requires

$$n \geq \frac{\log((1-\alpha)/\alpha)}{-\log(1-p_{\min,G_0})}.$$

When $p_{\min,G_0} \leq 1/2$ and $\alpha < 1/4$, the right-hand side is $\Omega(p_{\min,G_0}^{-1}\log(1/\alpha))$; near the saturated endpoint $p_{\min,G_0} \uparrow 1$ the exact expression must be used. If instead $q_\star \in [\eta, 1-\eta]$ and $0 < \Delta \leq \eta/2$, testing edge probabilities $q_\star - \Delta$ and $q_\star + \Delta$ has matching complexity $\Theta_\eta(\Delta^{-2}\log(1/\alpha))$.

Proof. For an edge $e \in G_0$, the probability that it is absent from its first n_{disc} row samples is

$$(1-P(e))^{n_{\mathrm{disc}}} \leq e^{-n_{\mathrm{disc}} p_{\min,G_0}} \leq \alpha/m.$$

A union bound shows that every witness edge is observed. Observed edges are nominally positive and, since $q_\star = 0$, belong to the robust core. The observed graph contains the weakly communicating graph G_0, and weak communication is preserved by adding edges, proving validity.

For sharpness, the defining pair reduces the certificate to testing a Bernoulli edge with $H_0 : q = 0$ against $H_1 : q = p$. Let r be the probability that a possibly randomized test declares the edge present on the all-zero sample. Type-I error at $q = 0$ is r, so $r \leq \alpha$. Under $q = p$, the all-zero sample has probability $(1-p)^n$, and the type-II error is at least $(1-p)^n(1-r) \geq (1-p)^n(1-\alpha)$. Requiring this to be at most α gives the displayed lower bound. The observe-once test attains the corresponding upper order.

For the interior claim, threshold the empirical Bernoulli mean at $q_\star$; Hoeffding's inequality gives $O(\Delta^{-2}\log(1/\alpha))$ samples. Conversely, uniformly for $q_\star \in [\eta, 1-\eta]$ and $0 < \Delta \leq \eta/2$,

$$\mathrm{kl}(q_\star - \Delta \,\|\, q_\star + \Delta) \leq C_\eta \Delta^2.$$

The Bretagnolle–Huber testing inequality then forces $n \geq c_\eta \Delta^{-2}\log(1/\alpha)$ whenever both errors are at most $\alpha < 1/4$. This proves both regimes. □

Corollary 3. *Assume $q_\star = 0$ and $C(P) = \mathcal{S}$. Let G_0 be a weakly communicating witness subgraph with m edges, let r_0 be the number of rows containing those edges, and let each witness edge have probability at least $p_{\min,G_0}$. Run π_0 and stop as soon as the observed-edge graph is weakly communicating. Let*

$$n_{\mathrm{disc}} := \left\lceil p_{\min,G_0}^{-1} \log(2m/\alpha) \right\rceil, \qquad M_{\mathrm{disc}} := r_0 n_{\mathrm{disc}}.$$

Then, with probability at least $1-\alpha$,

$$\tau_{\mathrm{topo}} \leq 2D_0(P) M_{\mathrm{disc}} \left\lceil \log_2 \frac{2M_{\mathrm{disc}}}{\alpha} \right\rceil.$$

Consequently the row-sample dependence is $\widetilde{O}(p_{\min,G_0}^{-1})$, and Theorem 7 shows that this dependence is sharp for one-edge-decisive certificates.

Proof. List each witness row n_{disc} times. The conditional expected waiting time for every requested visit is at most $D_0(P)$. Applying the geometric-block estimate (28) to each of the M_{disc} visits and taking a union bound shows that all requested row samples are collected by the displayed time except on an event $\mathcal{E}_{\mathrm{wait}}$ of probability at most $\alpha/2$. Independently of the waiting times, each collected row sample of a witness edge is, conditionally on the history, a draw from the true row (as in the proof of Lemma 12); the union bound in Theorem 7 therefore shows that, outside an event $\mathcal{E}_{\mathrm{abs}}$ of probability at most $\alpha/2$, every witness edge has appeared within its first n_{disc} samples. On $\mathcal{E}_{\mathrm{wait}}^c \cap \mathcal{E}_{\mathrm{abs}}^c$ the observed graph contains G_0 and is a valid common-core certificate. Since the stopping rule checks after every transition, it has stopped by that time. □

D.9 Occupation-weighted regret decomposition

Theorem 8. *Suppose $\delta < \delta_f^*(P)$ and $\mathcal{E}_{\mathrm{conf}}$ holds. In every episode beginning after τ_{topo}, let $\widetilde{Q}_k^* \in \mathcal{U}^P$ attain the robust gain of $\bar{\pi}_k$, let $\mu_k \in \Delta(\mathcal{S})$ satisfy $\mu_k^\top (\widetilde{Q}_k^*)_{\bar{\pi}_k} = \mu_k^\top$, with $\rho_k(s,a) := \mu_k(s)\mathbf{1}\{a = \bar{\pi}_k(s)\}$. If $g^{\mathrm{rob}}(P) \le g_k$, then*

$$\mathrm{Reg}_{\mathrm{pol}}(T) \le \tau_{\mathrm{topo}} \wedge T + \sum_{k: t_k \ge \tau_{\mathrm{topo}}} L_k(T) \sum_{s,a} \rho_k(s,a)\big(b_k(s,a) + \zeta_k(s)\big).$$

Here $\zeta_k(s)$ is interpreted only at the action selected by $\bar{\pi}_k$, as ρ_k is supported on that policy.

Theorem 8 is the general decomposition for the optimistic template: it holds for any episodic rule whose policies are completions supplied by Lemma 14 and whose optimistic gains satisfy $g^{\mathrm{rob}}(P) \le g_k$, and it carries no coverage coefficient, span, or condition number. Algorithm 2 is the instance of that template obtained by replacing the row-wise optimistic operator $\mathcal{L}_k$, after its exploration phase has certified $B_k = C(P)$, with a single optimistic center $Q_k \in \prod_{s \in B_k, a} \mathcal{C}^\circ_{k,B_k}(s,a)$, which turns $b_k(s,a)$ into $\sigma_\delta^f((Q_k)_{s,a}, h_k) - \sigma_\delta^f(P_{R,s,a}, h_k)$ and makes ζ_k vanish on the recurrent class of $\bar{\pi}_k$, since $\bar{\pi}_k$ agrees there with a greedy selector for Q_k. The proof of Theorem 12 carries out exactly this specialization, bounding the resulting b_k by $2H_{\mathrm{loc}}K_{\mathrm{face}}\beta_{\bar{\alpha}}$ and the occupation weights by C_{occ}. That proof establishes the specialization directly because the no-creation completion of Proposition 7 is stronger than the one Lemma 14 supplies.

Proof. We first record that the minimizing kernel $\widetilde{Q}_k^*$ exists. The set $\mathcal{U}^P$ is compact, and by Lemma 14 the policy $\bar{\pi}_k$ is unichain under every $\widetilde{Q} \in \mathcal{U}^P$, so $\mathrm{rank}(I - \widetilde{Q}_{\bar{\pi}_k}) = S - 1$ is constant there. Cramer's rule then makes $\widetilde{Q} \mapsto \mu_{\widetilde{Q}}^{\bar{\pi}_k}$, and hence $\widetilde{Q} \mapsto g_{\widetilde{Q}}^{\bar{\pi}_k}$, continuous on $\mathcal{U}^P$, so the infimum defining $g_{\mathrm{rob},P}^{\bar{\pi}_k}$ is attained.

The robust optimal gain and every robust policy gain lie in $[0,1]$ because all one-step rewards lie in $[0,1]$. Hence, for every episode,

$$g^{\mathrm{rob}}(P) - g_{\mathrm{rob},P}^{\bar{\pi}_k}(S_{t_k}) \le 1.$$

The algorithm ends the exploration phase and begins the first exploitation episode at τ_{topo}. Therefore the total length of all pre-certification portions before horizon T is at most $\tau_{\mathrm{topo}} \wedge T$, and their total policy-regret contribution is bounded by the same quantity.

Fix an exploitation episode k with $t_k \ge \tau_{\mathrm{topo}}$. Lemma 14 constructs $\bar{\pi}_k$ to be unichain in $\mathcal{G}_\wedge(P)$ and in every transition graph containing $\mathcal{G}_\wedge(P)$. By definition of the robust core,

$$\mathcal{G}_\wedge(P) \subseteq \mathcal{G}(\widetilde{Q}) \qquad \text{for every } \widetilde{Q} \in \mathcal{U}^P.$$

Thus $\bar{\pi}_k$ is unichain under every feasible adversarial kernel, in particular under the minimizing kernel $\widetilde{Q}_k^*$. Let μ_k be the unique stationary distribution of $(\widetilde{Q}_k^*)_{\bar{\pi}_k}$ and define its stationary occupation measure by

$$\rho_k(s,a) := \mu_k(s)\mathbf{1}\{a = \bar{\pi}_k(s)\}.$$

The optimistic eigenpair satisfies

$$g_k \mathbf{1} + h_k = \mathcal{L}_k h_k.$$

By the definition of the completion loss,

$$[\mathcal{L}_k h_k](s) = [\mathcal{L}_{k,\bar{\pi}_k} h_k](s) + \zeta_k(s).$$

Expanding the policy-specific optimistic operator gives

$$\begin{aligned}
[\mathcal{L}_{k,\bar{\pi}_k} h_k](s) &= r(s, \bar{\pi}_k(s)) + \sup_{q \in \mathcal{C}_k(s, \bar{\pi}_k(s))} \sigma_\delta^f(q, h_k) \\
&= r(s, \bar{\pi}_k(s)) + \sigma_\delta^f(P_{s,\bar{\pi}_k(s)}, h_k) + b_k(s, \bar{\pi}_k(s)) \\
&= [\mathcal{L}_{P,\bar{\pi}_k}^{\mathrm{rob}} h_k](s) + b_k(s, \bar{\pi}_k(s)).
\end{aligned}$$

Substituting the definitions of $\mathcal{L}_k$, ζ_k, and b_k gives the exact statewise identity

$$g_k + h_k(s) = [\mathcal{L}^{\mathrm{rob}}_{P,\bar{\pi}_k} h_k](s) + b_k(s, \bar{\pi}_k(s)) + \zeta_k(s). \tag{33}$$

The selected row of $\widetilde{Q}^*_k$ belongs to the corresponding ambiguity set. Hence

$$\sigma^f_\delta(P_{s,\bar{\pi}_k(s)}, h_k) = \min_{p \in \mathcal{U}^{P_{s,\bar{\pi}_k(s)}}} p^\top h_k \tag{34}$$

$$\leq (\widetilde{Q}^*_k)^\top_{s,\bar{\pi}_k(s)} h_k. \tag{35}$$

Substituting inequality (34) into equation (33) gives, for every state s,

$$\begin{aligned} g_k + h_k(s) \leq r(s, \bar{\pi}_k(s)) + (\widetilde{Q}^*_k)^\top_{s,\bar{\pi}_k(s)} h_k \\ + b_k(s, \bar{\pi}_k(s)) + \zeta_k(s). \end{aligned} \tag{36}$$

Multiply inequality (36) by $\mu_k(s)$ and sum over s. Because $\sum_s \mu_k(s) = 1$, the left side is

$$g_k + \sum_s \mu_k(s) h_k(s).$$

Stationarity gives

$$\begin{aligned} \sum_s \mu_k(s) (\widetilde{Q}^*_k)^\top_{s,\bar{\pi}_k(s)} h_k &= \sum_{s,s'} \mu_k(s) \widetilde{Q}^*_k(s' \mid s, \bar{\pi}_k(s)) h_k(s') \\ &= \sum_{s'} \mu_k(s') h_k(s'). \end{aligned}$$

Thus the bias terms on the two sides cancel, leaving

$$\begin{aligned} g_k \leq \sum_s \mu_k(s) r(s, \bar{\pi}_k(s)) \\ + \sum_s \mu_k(s) \big(b_k(s, \bar{\pi}_k(s)) + \zeta_k(s) \big). \end{aligned}$$

Since the chain is unichain, its average reward from every initial state is the stationary reward average:

$$g^{\bar{\pi}_k}_{\widetilde{Q}^*_k} = \sum_s \mu_k(s) r(s, \bar{\pi}_k(s)).$$

The kernel $\widetilde{Q}^*_k$ attains the infimum defining the robust policy gain, so

$$g^{\bar{\pi}_k}_{\widetilde{Q}^*_k} = g^{\bar{\pi}_k}_{\mathrm{rob},P}.$$

Using the definition of ρ_k, we obtain

$$g_k - g^{\bar{\pi}_k}_{\mathrm{rob},P} \leq \sum_{s,a} \rho_k(s,a) \big(b_k(s,a) + \zeta_k(s) \big).$$

The assumed optimism $g^{\mathrm{rob}}(P) \leq g_k$ now gives the episodewise bound

$$g^{\mathrm{rob}}(P) - g^{\bar{\pi}_k}_{\mathrm{rob},P} \leq \sum_{s,a} \rho_k(s,a) \big(b_k(s,a) + \zeta_k(s) \big).$$

Multiply by the elapsed episode length $L_k(T) = ((t_{k+1} \wedge T) - t_k)_+$ and sum over the episodes with $\tau_{\mathrm{topo}} \leq t_k < T$. Writing $\tau_k := t_{k+1} - t_k$, note $L_k(T) \leq \tau_k$ with equality for every episode that closes before T, so the truncation of the final episode is accounted for exactly. Adding the pre-certification contribution $\tau_{\mathrm{topo}} \wedge T$ proves the stated regret decomposition. □

D.10 A COMPLETE PLUG-IN GUARANTEE

Lemma 15. *For every $f \in \mathcal{F}$ and $\delta > 0$, $\varpi_{f,\delta}^{P_R}(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$ and*

$$|\sigma_\delta^f(q,u) - \sigma_\delta^f(q',u)| \le \mathrm{sp}(u)\varpi_{f,\delta}^{P_R}(\|q-q'\|_1)$$

whenever q, q' are corresponding rows of kernels in $\mathcal{K}(P_R)$. If $\delta < \delta_f^{\mathrm{bell}}(P_R)$, then $H_{\mathrm{loc}}(P_R) < +\infty$, and every normalized accumulation point of $v_{Q,\gamma}$ as $\gamma \uparrow 1$ has span at most $H_{\mathrm{loc}}(P_R)$, uniformly over $Q \in \mathcal{K}(P_R)$. On the nominal support faces,

$$|\sigma_\delta^f(q,u) - \sigma_\delta^f(q',u)| \le \mathrm{sp}(u)\varpi_{f,\delta,\mathrm{face}}^{P_R}(\|q-q'\|_1),$$
$$K_{\mathrm{face}}(P_R) < +\infty,$$

whenever q, q' are corresponding rows of kernels in $\mathcal{K}_{\mathrm{face}}(P_R)$.

Proof. We first prove joint continuity of the robust row functional and then prove the uniform bias bound by contradiction. Let

$$(q_n, u_n) \longrightarrow (q,u) \quad \text{in} \quad \Delta(R(P)) \times \mathbb{R}^{R(P)}.$$

For every n, compactness of $\mathcal{U}^{q_n}(\delta)$ gives a minimizing row p_n such that

$$D_f(p_n\|q_n) \le \delta, \qquad \sigma_\delta^f(q_n,u_n) = p_n^\top u_n.$$

The simplex is compact. Passing to a subsequence if necessary, there is $p \in \Delta(R(P))$ such that $p_n \to p$. The closed perspective representation of the f-divergence is lower semicontinuous in (p,q). Consequently,

$$D_f(p\|q) \le \liminf_{n\to\infty} D_f(p_n\|q_n) \le \delta.$$

Thus $p \in \mathcal{U}^q(\delta)$. Along the selected subsequence,

$$\liminf_{n\to\infty} \sigma_\delta^f(q_n,u_n) = \lim_{n\to\infty} p_n^\top u_n \tag{37}$$
$$= p^\top u \tag{38}$$
$$\ge \min_{\bar p \in \mathcal{U}^q(\delta)} \bar p^\top u \tag{39}$$
$$= \sigma_\delta^f(q,u). \tag{40}$$

Every subsequence attaining the lower limit has a further convergent subsequence to which equation (37) applies. Therefore

$$\liminf_{n\to\infty} \sigma_\delta^f(q_n,u_n) \ge \sigma_\delta^f(q,u).$$

For the reverse inequality, fix a minimizer p^* at (q,u). For $\tau \in (0,1)$ define

$$p^\tau := (1-\tau)p^* + \tau q.$$

Joint convexity of the f-divergence in its first argument, with the second argument fixed at q, gives

$$\begin{aligned} D_f(p^\tau\|q) &\le (1-\tau)D_f(p^*\|q) + \tau D_f(q\|q) \\ &\le (1-\tau)\delta \\ &< \delta. \end{aligned}$$

For the perturbed nominal row set

$$p_n^\tau := (1-\tau)p^* + \tau q_n.$$

Then $p_n^\tau \to p^\tau$. We verify directly that

$$D_f(p_n^\tau\|q_n) \longrightarrow D_f(p^\tau\|q).$$

Fix a coordinate i. There are three cases.

If $q_i > 0$, then $q_{n,i} > 0$ eventually and

$$q_{n,i} f\left(\frac{p_{n,i}^\tau}{q_{n,i}}\right) \longrightarrow q_i f\left(\frac{p_i^\tau}{q_i}\right)$$

by continuity of the finite convex function f on $(0,\infty)$.

If $q_i = 0$ and $p_i^* > 0$, feasibility of p^* implies $f'(\infty) < +\infty$. Moreover,

$$p_{n,i}^\tau = (1-\tau)p_i^* + \tau q_{n,i}.$$

When $q_{n,i} > 0$,

$$\begin{aligned} &q_{n,i} f\left(\frac{(1-\tau)p_i^* + \tau q_{n,i}}{q_{n,i}}\right) \\ &\quad = \big((1-\tau)p_i^* + \tau q_{n,i}\big)\frac{f(((1-\tau)p_i^* + \tau q_{n,i})/q_{n,i})}{((1-\tau)p_i^* + \tau q_{n,i})/q_{n,i}}. \end{aligned}$$

The first factor converges to $(1-\tau)p_i^*$ and the argument in the last fraction tends to $+\infty$. Hence the coordinate contribution converges to

$$(1-\tau)p_i^* f'(\infty),$$

which is exactly the singular-coordinate contribution to $D_f(p^\tau\|q)$. The same conclusion is immediate along indices for which $q_{n,i} = 0$ from the singular-mass convention.

Finally, if $q_i = p_i^* = 0$, then $p_{n,i}^\tau = \tau q_{n,i}$ and the coordinate contribution is

$$q_{n,i} f(\tau) \longrightarrow 0.$$

This equals the $0 \cdot f(0/0) = 0$ contribution at the limit. There are only finitely many coordinates, so summing the three coordinatewise limits proves the claimed divergence convergence.

Since $D_f(p^\tau\|q) < \delta$, the convergence implies that $D_f(p_n^\tau\|q_n) \le \delta$ for all sufficiently large n. Therefore

$$\begin{aligned} \limsup_{n\to\infty} \sigma_\delta^f(q_n, u_n) &\le \lim_{n\to\infty} (p_n^\tau)^\top u_n \\ &= (p^\tau)^\top u \\ &= (1-\tau)(p^*)^\top u + \tau q^\top u. \end{aligned}$$

Letting $\tau \downarrow 0$ gives

$$\limsup_{n\to\infty} \sigma_\delta^f(q_n, u_n) \le (p^*)^\top u = \sigma_\delta^f(q, u).$$

Together with the lower bound, this proves joint continuity of $(q,u) \mapsto \sigma_\delta^f(q,u)$.

We next prove that the local modulus tends to zero. The set $\mathcal{K}(P_R)$ is compact, because it is a closed subset of the finite product of simplices. The normalized bias set

$$\mathfrak{U}_1 := \{u \in \mathbb{R}^{R(P)} : \min_x u(x) = 0,\ \mathrm{sp}(u) \le 1\}$$

is also compact: every coordinate of $u \in \mathfrak{U}_1$ lies in $[0,1]$. The joint continuity proved above is uniform on the compact set of all triples $(Q_{s,a}, Q'_{s,a}, u)$ with $Q, Q' \in \mathcal{K}(P_R)$ and $u \in \mathfrak{U}_1$. It follows that

$$\varpi_{f,\delta}^{P_R}(\epsilon) \longrightarrow 0 \qquad \text{as } \epsilon \downarrow 0.$$

The modulus is nondecreasing in ϵ, so the convergence may be written $\varpi_{f,\delta}^{P_R}(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$.

Now fix arbitrary $u \in \mathbb{R}^{R(P)}$. If $\mathrm{sp}(u) = 0$, then $u = c\mathbf{1}$ and additive homogeneity gives

$$\sigma_\delta^f(q,u) = c = \sigma_\delta^f(q',u),$$

so the asserted inequality is immediate. If $c_u := \mathrm{sp}(u) > 0$, let

$$\bar{u} := \frac{u - (\min_x u(x))\mathbf{1}}{c_u}.$$

Then $\min_x \bar{u}(x) = 0$ and $\mathrm{sp}(\bar{u}) = 1$. Additive and positive homogeneity give

$$\begin{aligned} \sigma_\delta^f(q,u) &= \min_x u(x) + c_u \sigma_\delta^f(q,\bar{u}), \\ \sigma_\delta^f(q',u) &= \min_x u(x) + c_u \sigma_\delta^f(q',\bar{u}). \end{aligned}$$

Subtracting and applying the definition of the local modulus yields

$$|\sigma_\delta^f(q,u) - \sigma_\delta^f(q',u)| \le \mathrm{sp}(u)\varpi_{f,\delta}^{P_R}(\|q-q'\|_1).$$

We now establish a uniform Lipschitz bound inside each nominal support face. Fix a row and let F be the face supported on its nominal support. For $u \in \mathfrak{U}_1$, the convex function $T_u(q) := \sigma_\delta^f(q,u)$ takes values in $[0,1]$ on F. Every row of a kernel in $\mathcal{K}_{\mathrm{face}}(P_R)$ has each nominally positive coordinate at least $p_{\min,R}(P)/2$: indeed the total-variation change of one coordinate is at most half the ℓ_1 distance, and $\bar{\epsilon}_{\mathrm{face}}(P_R) \le p_{\min,R}(P)/4$. Thus these rows lie a common positive distance from the relative boundary of F.

The elementary interior estimate for a convex function bounded in an interval of length one now gives, uniformly in u,

$$|T_u(q) - T_u(q')| \le \frac{2}{p_{\min,R}(P)}\|q-q'\|_1.$$

For completeness, apply the one-dimensional secant-slope inequality to $t \mapsto T_u(q+tz)$ for tangent directions z with $\|z\|_1 = 1$; the coordinate lower bound permits a segment of length $p_{\min,R}(P)/2$ in both directions, while the function oscillates by at most one. Integrating the resulting directional-slope bound along the segment from q to q' proves the display. Positive homogeneity gives the same result for arbitrary u, multiplied by $\mathrm{sp}(u)$. Maximizing over the finitely many rows proves

$$K_{\mathrm{face}}(P_R) \le \frac{2}{p_{\min,R}(P)} < +\infty.$$

We next record the scalar-Bellman stability built into the local class. By the definition of $\epsilon_{\mathrm{wc}}(P_R)$, every $Q \in \mathcal{K}(P_R)$ satisfies $\delta < \delta_f^{\mathrm{bell}}(Q)$. In the no-creation class this follows from continuity of the finitely many deletion costs on the fixed nominal support face; in the creation class $\delta_f^{\mathrm{bell}} \equiv +\infty$.

It remains to prove uniform boundedness of the discounted bias spans. Suppose for contradiction that $H_{\mathrm{loc}}(P_R) = +\infty$. Then there are $Q_n \in \mathcal{K}(P_R)$ and $\gamma_n \in (0,1)$ such that, with $v_n := v_{Q_n,\gamma_n}$,

$$L_n := \mathrm{sp}(v_n) \longrightarrow +\infty.$$

The fixed-point equation and the elementary bound $\mathrm{sp}(\mathcal{L}_{Q,\gamma}^{\mathrm{rob}} u) \le 1 + \gamma\,\mathrm{sp}(u)$ give

$$L_n \le 1 + \gamma_n L_n, \qquad L_n \le \frac{1}{1-\gamma_n}. \tag{41}$$

If a subsequence of γ_n were bounded above by $1-\varepsilon$ for some $\varepsilon > 0$, inequality (41) would give $L_n \le 1/\varepsilon$ on that subsequence, a contradiction. Hence

$$\gamma_n \longrightarrow 1.$$

Fix $s_\circ \in R(P)$ and define

$$h_n := v_n - v_n(s_\circ)\mathbf{1}, \qquad y_n := \frac{h_n}{L_n}.$$

Then $y_n(s_\circ) = 0$ and $\mathrm{sp}(y_n) = 1$. Consequently every coordinate of y_n lies in $[-1,1]$. Compactness gives a subsequence, not relabeled, such that

$$Q_n \longrightarrow Q_\infty \in \mathcal{K}(P_R), \qquad y_n \longrightarrow y, \qquad \mathrm{sp}(y) = 1.$$

Let

$$c_n := v_n(s_\circ), \qquad \lambda_n := (1-\gamma_n)c_n.$$

Using additive homogeneity in the discounted fixed-point equation gives, for every state s,

$$h_n(s) + \lambda_n = \max_a\{r(s,a) + \gamma_n \sigma_\delta^f(Q_{n,s,a}, h_n)\}.$$

Divide by L_n and use positive homogeneity:

$$y_n(s) + \frac{\lambda_n}{L_n} = \max_a\left\{\frac{r(s,a)}{L_n} + \gamma_n\sigma_\delta^f(Q_{n,s,a}, y_n)\right\}.$$

Because $y_n \in [-1,1]^{R(P)}$, every robust expectation on the right lies in $[-1,1]$. Hence λ_n/L_n is bounded. Pass to a further subsequence on which $\lambda_n/L_n \to \lambda$. Joint continuity of the row functional, $L_n \to \infty$, and $\gamma_n \to 1$ yield

$$y(s) + \lambda = \max_a \sigma_\delta^f(Q_{\infty,s,a}, y) \qquad (s \in R(P)).$$

Let $m := \min_s y(s)$ and $M := \max_s y(s)$. At a maximizing state, every probability expectation is at most M, so

$$M + \lambda \le M, \qquad \text{hence } \lambda \le 0.$$

At a minimizing state, every probability expectation is at least m, so

$$m + \lambda \ge m, \qquad \text{hence } \lambda \ge 0.$$

Thus $\lambda = 0$, and y is a fixed point of the recession operator

$$y(s) = \max_a \sigma_\delta^f(Q_{\infty,s,a}, y).$$

Since $Q_\infty \in \mathcal{K}(P_R)$, we have $\delta < \delta_f^{\mathrm{bell}}(Q_\infty)$. The recession fixed-point argument in the proof of Theorem 3 shows that every such recession fixed point is constant. This contradicts $\mathrm{sp}(y) = 1$. Therefore $H_{\mathrm{loc}}(P_R) < +\infty$.

Finally, every normalized vanishing-discount accumulation point is the limit of functions whose spans are bounded by $H_{\mathrm{loc}}(P_R)$. Continuity of the span under coordinatewise convergence gives the same bound for the limit. This completes the proof. □

Lemma 16. *Let $Q \in \mathcal{K}(P_R)$, let $(g_Q, h_Q) \in \mathrm{Eig}(\mathcal{L}_Q^{\mathrm{rob}})$ be obtained as a vanishing-discount accumulation point, and let $\pi_Q \in \mathrm{Gr}_Q(h_Q)$. If*

$$\|Q - P_R\|_{\max,1} \le \eta,$$

then, for every $x \in R(P)$,

$$g^{\mathrm{rob}}(P_R) - g_{\mathrm{rob},P_R}^{\pi_Q}(x) \le 2H_{\mathrm{loc}}(P_R)\varpi_{f,\delta}^{P_R}(\eta).$$

The policy π_Q need not be unichain under P_R, under Q, or simultaneously over either ambiguity set.

Proof. Write

$$H := \mathrm{sp}(h_Q), \qquad e := H\varpi_{f,\delta}^{P_R}(\eta).$$

Lemma 15 and the definition of η imply, simultaneously for every state and action,

$$|\sigma_\delta^f(P_{R,s,a}, h_Q) - \sigma_\delta^f(Q_{s,a}, h_Q)| \le e. \tag{42}$$

The Bellman equation for Q is

$$g_Q + h_Q(s) = \max_a\{r(s,a) + \sigma_\delta^f(Q_{s,a}, h_Q)\}.$$

Using the upper inequality in (42) inside the maximum gives

$$\begin{aligned}[\mathcal{L}_{P_R}^{\mathrm{rob}} h_Q](s) &= \max_a\{r(s,a) + \sigma_\delta^f(P_{R,s,a}, h_Q)\} \\ &\le \max_a\{r(s,a) + \sigma_\delta^f(Q_{s,a}, h_Q) + e\} \\ &= g_Q + h_Q(s) + e.\end{aligned}$$

We next derive the required gain upper bound directly. Fix an arbitrary stationary deterministic policy π. For every state s, choose a row $p_s^\pi \in \mathcal{U}^{P_{R,s,\pi(s)}}(\delta)$ attaining $\sigma_\delta^f(P_{R,s,\pi(s)}, h_Q)$, and combine these rows into a rectangular kernel $\widetilde{P}^\pi \in \mathcal{U}^{P_R}$. Since the selected action backup is bounded by the maximum action backup,

$$r(s,\pi(s)) + (\widetilde{P}^\pi)_{s,\pi(s)}^\top h_Q \le h_Q(s) + g_Q + e. \tag{43}$$

Run policy π under $\widetilde{P}^{\pi}$, condition on $S_0 = x$, and sum inequality (43) from $t = 0$ to $N-1$. The conditional transition law gives

$$\mathbb{E}^{\pi}_{\widetilde{P}^{\pi}}\left[\sum_{t=0}^{N-1} r(S_t, \pi(S_t)) \mid S_0 = x\right] \leq N(g_Q + e) + h_Q(x) - \mathbb{E}^{\pi}_{\widetilde{P}^{\pi}}[h_Q(S_N) \mid S_0 = x].$$

Divide by N. The last two terms have absolute difference at most H/N, so taking the limsup gives

$$g^{\pi}_{\widetilde{P}^{\pi}}(x) \leq g_Q + e.$$

Because $\widetilde{P}^{\pi} \in \mathcal{U}^{P_R}$,

$$g^{\pi}_{\mathrm{rob},P_R}(x) = \inf_{\widetilde{P} \in \mathcal{U}^{P_R}} g^{\pi}_{\widetilde{P}}(x) \leq g_Q + e.$$

Taking the supremum over π and applying Theorem 3 gives

$$g^{\mathrm{rob}}(P_R) \leq g_Q + e. \tag{44}$$

We now lower-bound the true robust gain of the greedy plug-in policy. Greediness gives the policywise equality

$$g_Q + h_Q(s) = r(s, \pi_Q(s)) + \sigma^f_{\delta}(Q_{s,\pi_Q(s)}, h_Q).$$

Fix an arbitrary $\widetilde{P} \in \mathcal{U}^{P_R}$. Feasibility of its selected row and the lower inequality in (42) give

$$\begin{aligned} r(s, \pi_Q(s)) + \widetilde{P}^{\top}_{s,\pi_Q(s)} h_Q & \tag{45}\\ \geq r(s, \pi_Q(s)) + \sigma^f_{\delta}(P_{R,s,\pi_Q(s)}, h_Q) & \tag{46}\\ \geq r(s, \pi_Q(s)) + \sigma^f_{\delta}(Q_{s,\pi_Q(s)}, h_Q) - e & \tag{47}\\ = g_Q + h_Q(s) - e. & \tag{48} \end{aligned}$$

Sum inequality (45) along the trajectory under $(\widetilde{P}, \pi_Q)$ for N steps. We obtain

$$\mathbb{E}^{\pi_Q}_{\widetilde{P}}\left[\sum_{t=0}^{N-1} r(S_t, \pi_Q(S_t)) \mid S_0 = x\right] \geq N(g_Q - e) + h_Q(x) - \mathbb{E}^{\pi_Q}_{\widetilde{P}}[h_Q(S_N) \mid S_0 = x].$$

After division by N, the bias endpoint term again vanishes. Therefore

$$g^{\pi_Q}_{\widetilde{P}}(x) \geq g_Q - e.$$

Since $\widetilde{P} \in \mathcal{U}^{P_R}$ was arbitrary,

$$g^{\pi_Q}_{\mathrm{rob},P_R}(x) \geq g_Q - e. \tag{49}$$

Combining inequalities (44) and inequality (49) yields

$$\begin{aligned} g^{\mathrm{rob}}(P_R) - g^{\pi_Q}_{\mathrm{rob},P_R}(x) &\leq (g_Q + e) - (g_Q - e)\\ &= 2e\\ &= 2\,\mathrm{sp}(h_Q)\varpi^{P_R}_{f,\delta}(\eta)\\ &\leq 2H_{\mathrm{loc}}(P_R)\varpi^{P_R}_{f,\delta}(\eta). \end{aligned}$$

At no point did we use the number of recurrent classes of π_Q. □

Lemma 17. *Suppose $\mathcal{G}(P) \in \mathcal{M}$ and either $f \in \mathcal{F}_{\mathrm{ne}}$ or $C(P) = \mathcal{S}$. With probability at least $1 - 2\alpha/3$, the first m_T steps contain at least $n_T(P,\alpha)$ samples from every $(s,a) \in R(P) \times \mathcal{A}$. If additionally $n_T(P,\alpha) \geq n_{\mathrm{supp}}(P,\alpha)$, then the observed support of every relevant row equals the support of P_R. If also $f \in \mathcal{F}_{\mathrm{ne}}$, then*

$$S_{m_T} \in C(P), \qquad \widehat{R}_T = C(P)$$

on the same event. Moreover

$$n_T(P,\alpha) = \widetilde{\Omega}\left(\frac{T^{2/3}}{D_0(P) d_R A}\right).$$

Proof. Write $R := R(P)$ and $m_R := d_R A$.

If $n_T(P, \alpha) = 0$, the sample-count assertion is immediate. We therefore bound the time required to collect any prescribed $n \geq 1$ samples per relevant pair. Fix any target pair $(s, a) \in R \times \mathcal{A}$. The global definition of $D_0(P)$ gives, from every $x \in \mathcal{S}$,

$$\mathbb{E}_x^{\pi_0}[\tau_{s,a}] \leq D_0(P) - 1 < D_0(P).$$

As above, a request is complete only after the target action's successor is observed, so its conditional expected completion time is at most $D_0(P)$. Inequality (28), applied conditionally at the beginning of each request with $D = D_0(P)$, shows that a requested visit to (s, a) is not completed within b blocks of length $2D_0(P)$ with probability at most 2^{-b}.

List every pair in $R \times \mathcal{A}$ exactly n times, in any fixed order. This gives $m_R n$ requested visits. Complete these requests sequentially: after one requested pair is visited, begin waiting for the next pair on the list. Set

$$b_n := \left\lceil \log_2 \frac{6 m_R n}{\alpha} \right\rceil.$$

For each request the conditional failure probability within $2D_0(P)b_n$ steps is at most 2^{-b_n}. A union bound over all $m_R n$ requests gives

$$\begin{aligned} \mathbb{P}^{\pi_0}(\text{some request fails}) &\leq m_R n \, 2^{-b_n} \\ &\leq \frac{\alpha}{6}. \end{aligned}$$

Thus, with probability at least $1 - \alpha/6$, all requested visits are completed from an arbitrary initial state within

$$2D_0(P) m_R n \left\lceil \log_2 \frac{6 m_R n}{\alpha} \right\rceil = \mathfrak{t}_P(n, \alpha)$$

steps. Taking $n = n_T(P, \alpha)$ proves the sample-count assertion, with room in the stated probability for support recovery.

We now prove that every positive relevant edge is observed. Each completed request to a pair (s, a) observes a successor drawn from $P_{s,a}$ conditionally on the history preceding that request. No conditioning on the realized visit times is used, and no independence across different pairs is required. If $P(s' \mid s, a) = p > 0$, iterated conditioning shows that the edge $s \to s'$ is absent from the observed successors of the first n completed requests to its row with probability at most

$$(1 - p)^n \leq e^{-np}.$$

For $s, s' \in R(P)$, every positive edge has $p \geq p_{\min,R}(P)$. There are at most $d_R^2 A$ candidate triples. Hence a union bound gives

$$\mathbb{P}(\text{some positive edge of } P_R \text{ is missed}) \leq d_R^2 A e^{-n p_{\min,R}(P)}.$$

If $n \geq n_{\text{supp}}(P, \alpha)$, the right side is at most $\alpha/6$. Combining this event with the sample-collection event above gives probability at least $1 - \alpha/3$, which is stronger than $1 - 2\alpha/3$. On this event the observed support of every relevant row is exactly the nominal support.

Suppose now that $f \in \mathcal{F}_{\text{ne}}$. On the same event $S_{m_T} \in C(P)$. The union transition graph induced by all actions is strongly connected on $C(P)$ by Lemma 6(iii). Every one of its positive edges has been observed, so all states in $C(P)$ lie in the same observed strongly connected component as S_{m_T}. Conversely, no nominal edge leaves $C(P)$ under any action, since $C(P)$ is closed in $\mathcal{G}(P)$ by Lemma 6(i). Thus no state outside $C(P)$ can belong to the strongly connected component containing S_{m_T}. Therefore

$$\widehat{R}_T = C(P).$$

Finally, we invert the elapsed-time bound. Let $m := m_T$ and, for all sufficiently large m, define

$$\widetilde{n}_m := \left\lfloor \frac{m}{4D_0(P) m_R \lceil \log_2(6 m_R m/\alpha) \rceil} \right\rfloor.$$

Because $\widetilde{n}_m \le m$, we have

$$\left\lceil \log_2 \frac{6m_R\widetilde{n}_m}{\alpha} \right\rceil \le \left\lceil \log_2 \frac{6m_R m}{\alpha} \right\rceil.$$

The pair-collection term at $n = \widetilde{n}_m$ is therefore at most

$$\begin{aligned} &2D_0(P)m_R\widetilde{n}_m \left\lceil \log_2 \frac{6m_R\widetilde{n}_m}{\alpha} \right\rceil \\ &\quad \le 2D_0(P)m_R \frac{m}{4D_0(P)m_R \lceil \log_2(6m_R m/\alpha) \rceil} \left\lceil \log_2 \frac{6m_R m}{\alpha} \right\rceil \\ &\quad = \frac{m}{2}. \end{aligned}$$

Hence $\mathfrak{t}_P(\widetilde{n}_m, \alpha) \le m/2 \le m$, and the maximality in the definition of $n_T(P, \alpha)$ gives

$$n_T(P, \alpha) \ge \widetilde{n}_{m_T}.$$

Since $m_T = \lceil T^{2/3} \rceil$, this proves

$$n_T(P, \alpha) = \widetilde{\Omega}\left(\frac{T^{2/3}}{D_0(P) d_R A} \right).$$

□

Lemma 18. *Let $f \in \mathcal{F}_{\mathrm{ne}}$, $\mathcal{G}(P) \in \mathcal{M}$ and $\delta < \delta_f^{\mathrm{bell}}(P)$. Then $\delta < \delta_f^{\mathrm{bell}}(P_R)$ and, for every $x \in C(P)$ and every $\pi \in \Pi$,*

$$g^{\pi}_{\mathrm{rob},P}(x) = g^{\pi|_{C(P)}}_{\mathrm{rob},P_R}(x), \qquad g^{\mathrm{rob}}(P) = g^{\mathrm{rob}}(P_R).$$

If moreover $\delta < \delta_f^*(P)$, then $\delta < \delta_f^*(P_R)$.

Proof. Suppose, to the contrary, that $\delta \ge \delta_f^{\mathrm{bell}}(P_R)$. Then there are disjoint nonempty $I, J \subseteq C(P)$ such that I is δ-closed for P_R and J is closable in $\mathcal{G}(P_R)$. Since $C(P)$ is closed in $\mathcal{G}(P)$ by Lemma 6(i), the actions witnessing closability in $\mathcal{G}(P_R)$ also witness closability in $\mathcal{G}(P)$. The nominal mass outside I in every row based in I is unchanged by restriction, so $\rho^{P}_{f,\mathrm{all}}(I) = \rho^{P_R}_{f,\mathrm{all}}(I) \le \delta$. That would give $\delta \ge \delta_f^{\mathrm{bell}}(P)$, a contradiction.

If $\delta < \delta_f^*(P)$, then $\mathcal{G}_\wedge(P_R) = \mathcal{G}_\wedge(P)|_{C(P)}$: every nominal row based in $C(P)$ is supported in $C(P)$ and has the same deletion costs before and after restriction. The graph $\mathcal{G}_\wedge(P)$ is weakly communicating, its communicating class is contained in $C(P)$, and deleting states outside $C(P)$ removes only states transient under every deterministic policy. Thus $\mathcal{G}_\wedge(P_R)$ is weakly communicating on $C(P)$, which is equivalent to $\delta < \delta_f^*(P_R)$.

By Lemma 6(i), $C(P)$ is closed in $\mathcal{G}(P)$. Since $f \in \mathcal{F}_{\mathrm{ne}}$, Theorem 1 gives $\mathcal{G}(\widetilde{Q}) \subseteq \mathcal{G}(P)$ for every $\widetilde{Q} \in \mathcal{U}^P(\delta)$, so $C(P)$ is closed in $\mathcal{G}(\widetilde{Q})$ as well. Hence for $x \in C(P)$ every trajectory of $(\widetilde{Q}, \pi)$ started at x stays in $C(P)$ almost surely, and its reward stream depends only on the rows of $\widetilde{Q}$ based in $C(P)$.

For $s \in C(P)$ the nominal row $P_{s,a}$ is supported in $C(P)$, so the constraint $D_f(\cdot \| P_{s,a}) \le \delta$ describes the same set of distributions on $C(P)$ whether read in $\mathcal{U}^P(\delta)$ or in $\mathcal{U}^{P_R}(\delta)$. By rectangularity the two infima defining $g^{\pi}_{\mathrm{rob},P}(x)$ and $g^{\pi|_{C(P)}}_{\mathrm{rob},P_R}(x)$ therefore range over the same family of reward streams, proving the first equality in Lemma 18.

Both optimal gains are scalar by Theorem 3. Every $\pi \in \Pi$ restricts to a policy on $C(P)$ and every policy on $C(P)$ extends to $\mathcal{S}$, so taking suprema in the first equality of Lemma 18 at any fixed $x \in C(P)$ gives $g^{\mathrm{rob}}(P) = g^{\mathrm{rob}}(P_R)$. □

Algorithm 1 Explore–then–commit (ETC)

Input: T, α, δ, $s_\circ$; $m_T := \lceil T^{2/3} \rceil$
Play $A_t \sim \pi_0(\cdot \mid S_t)$ for $t < m_T$
Form the observed graph $\widehat{G}_{m_T}(s,a)$ of transitions seen from (s,a)
If $f \in \mathcal{F}_{\mathrm{ne}}$, set $\widehat{R}_T$ to the strongly connected component of $\bigcup_a \widehat{G}_{m_T}(\cdot,a)$ containing S_{m_T}; else $\widehat{R}_T := \mathcal{S}$
if some row $(s,a) \in \widehat{R}_T \times \mathcal{A}$ has no observations, or its restriction to $\widehat{R}_T$ is not stochastic **then**
    Play the fixed policy π_0 for $m_T \le t < T$ and stop
$\widehat{P}_T := \widehat{P}_{m_T}|_{\widehat{R}_T \times \mathcal{A}}$
if $\delta < \delta_f^{\mathrm{bell}}(\widehat{P}_T)$ **then**
    Fix a reference state $s_\star \in \widehat{R}_T$ (e.g. $s_\star := S_{m_T}$) and take $\widehat{h}_T$ as a $\gamma \uparrow 1$ accumulation point of $v_{\widehat{P}_T,\gamma} - v_{\widehat{P}_T,\gamma}(s_\star)\mathbf{1}$
    Play the policy that is greedy for $\widehat{h}_T$ on $\widehat{R}_T$ and equals π_0 on $\mathcal{S} \setminus \widehat{R}_T$, for $m_T \le t < T$
else
    Play an arbitrary fixed $\pi \in \Pi_{\mathrm{d}}$ for $m_T \le t < T$

Theorem 9. *Suppose*

$$\mathcal{G}(P) \in \mathcal{M}, \qquad \delta < \delta_f^{\mathrm{bell}}(P), \qquad \eta_T \le \bar{\epsilon}_{\mathrm{wc}}(P_R),$$
$$f \in \mathcal{F}_{\mathrm{ne}} \text{ or } C(P) = \mathcal{S},$$
$$f \in \mathcal{F}_{\mathrm{ne}} \implies n_T(P,\alpha) \ge n_{\mathrm{supp}}(P,\alpha).$$

Then Algorithm 1 satisfies, with probability at least $1-\alpha$*,*

$$\mathrm{Reg}_{\mathrm{pol}}(T) \le m_T + 2(T - m_T) H_{\mathrm{loc}}(P_R) \varpi_{f,\delta}^{P_R}(\eta_T).$$

For each fixed instance and fixed α*, the displayed deterministic upper bound is* $o(T)$*. Taking* $\alpha_T = T^{-2}$ *gives* $\mathbb{E}[\mathrm{Reg}_{\mathrm{pol}}(T)] = o(T)$*. This holds throughout* $\delta_f^*(P) \le \delta < \delta_f^{\mathrm{bell}}(P)$*. Below* $\delta_f^\dagger(P)$ *every ambiguity member is weakly communicating; above it, scalar robust control follows instead from the exact Bellman criterion. If moreover* $n_T(P,\alpha) \ge n_{\mathrm{supp}}(P,\alpha)$ *and* $\eta_T \le \bar{\epsilon}_{\mathrm{face}}(P_R)$*, then*

$$\mathrm{Reg}_{\mathrm{pol}}(T) = \widetilde{O}\Big(\big[1 + H_{\mathrm{loc}}(P_R) K_{\mathrm{face}}(P_R) \sqrt{D_0(P) d_R A [S + \log(SAT/\alpha)]}\big] T^{2/3} \Big).$$

Proof. We verify the restriction to $R(P)$, the high-probability event, the regret inequality, and the asymptotic conclusion separately.

We first reduce robust control to the relevant state set. If $f \notin \mathcal{F}_{\mathrm{ne}}$, Theorem 9 assumes $C(P) = \mathcal{S}$, so $R(P) = \mathcal{S}$ and there is nothing to prove. Suppose $f \in \mathcal{F}_{\mathrm{ne}}$. Then $f'(\infty) = +\infty$, so no feasible ambiguity row can place positive mass on a state to which the nominal row assigns zero mass. Because $C(P)$ is closed under every nominal action, every row centered at $s \in C(P)$ remains supported on $C(P)$. Thus, once the process enters $C(P)$, both the nominal dynamics and every ambiguity kernel remain in $C(P)$ under every policy. Rewards and all robust Bellman backups from that time onward depend only on the restricted kernel P_R.

We also need $\delta < \delta_f^{\mathrm{bell}}(P_R)$. This is precisely the first conclusion of Lemma 18; its proof extends any disjoint closed–closable obstruction on $C(P)$ to the original state space.

We next construct the joint good event. Let $\mathcal{E}_{\mathrm{exp}}$ be the event in Lemma 17. It has probability at least $1 - 2\alpha/3$. On this event every relevant row has at least $n_T(P,\alpha)$ samples. In the no-creation case, the assumed support threshold also gives $\widehat{R}_T = R(P)$. Whenever $n_T(P,\alpha) \ge n_{\mathrm{supp}}(P,\alpha)$, Lemma 17 also gives $\mathrm{supp}(\widehat{P}_{T,s,a}) = \mathrm{supp}(P_{R,s,a})$ for every relevant row.

Let $\mathcal{E}_{\mathrm{conf}}$ be the multinomial confidence event

$$\|\widehat{P}_{T,s,a} - P_{R,s,a}\|_1 \le \beta_{\alpha/3}(N_{m_T}(s,a)) \quad \text{for all } (s,a) \in R(P) \times \mathcal{A}.$$

Applying inequality (25) and the time-union calculation in the proof of Lemma 12, with failure probability $\alpha/3$, gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{conf}}) \ge 1 - \alpha/3.$$

The radius is nonincreasing in the relevant range. Therefore, on $\mathcal{E}_{\exp} \cap \mathcal{E}_{\mathrm{conf}}$,

$$\|\widehat{P}_T - P_R\|_{\max,1} \le \beta_{\alpha/3}(n_T(P,\alpha)) = \eta_T.$$

A union bound gives

$$\mathbb{P}(\mathcal{E}_{\exp} \cap \mathcal{E}_{\mathrm{conf}}) \ge 1 - \alpha.$$

Assume now that $\eta_T \le \bar{\epsilon}_{\mathrm{wc}}(P_R)$. Then $\widehat{P}_T \in \mathcal{K}(P_R)$ and, by the definition of $\epsilon_{\mathrm{wc}}(P_R)$,

$$\delta < \delta_f^{\mathrm{bell}}(\widehat{P}_T).$$

Thus the algorithm does not invoke its fallback. Theorem 3 supplies a vanishing-discount additive eigenpair for $\widehat{P}_T$ and makes its greedy selector well defined and robust optimal for the empirical kernel.

We now bound the regret accumulated in the two phases. Every reward and every average reward lies in $[0,1]$. Therefore the per-step robust policy gap during exploration is at most one. The first m_T steps contribute at most m_T to $\mathrm{Reg}_{\mathrm{pol}}(T)$.

During exploitation the algorithm uses the fixed policy $\widehat{\pi}_T$. Apply Lemma 16 with $Q = \widehat{P}_T$ and $\eta = \eta_T$. For every possible exploitation starting state $x \in R(P)$,

$$g^{\mathrm{rob}}(P_R) - g^{\widehat{\pi}_T}_{\mathrm{rob},P_R}(x) \le 2H_{\mathrm{loc}}(P_R)\varpi^{P_R}_{f,\delta}(\eta_T).$$

At each exploitation time this inequality may therefore be applied with $x = S_t$; its right-hand side is independent of the realized state. Summing the per-step policy-regret definition over the $T - m_T$ exploitation steps and adding the exploration loss give

$$\mathrm{Reg}_{\mathrm{pol}}(T) \le m_T + 2(T - m_T)H_{\mathrm{loc}}(P_R)\varpi^{P_R}_{f,\delta}(\eta_T).$$

For horizons at which either deterministic threshold in Theorem 9 is not met, rewards in $[0,1]$ always give the fallback bound $\mathrm{Reg}_{\mathrm{pol}}(T) \le T$.

We next prove sublinearity. By construction,

$$\frac{m_T}{T} \longrightarrow 0.$$

Lemma 17 gives $n_T(P,\alpha) \to \infty$. The definition of the confidence radius then gives

$$\eta_T = \beta_{\alpha/3}(n_T(P,\alpha)) \longrightarrow 0.$$

The support threshold and neighborhood threshold are finite constants of the fixed instance, so the applicable thresholds hold for every sufficiently large T. Lemma 15 gives

$$\varpi^{P_R}_{f,\delta}(\eta_T) \longrightarrow 0, \qquad H_{\mathrm{loc}}(P_R) < +\infty.$$

Dividing the regret inequality by T therefore yields

$$\begin{aligned}\frac{\mathrm{Reg}_{\mathrm{pol}}(T)}{T} &\le \frac{m_T}{T} + 2\left(1 - \frac{m_T}{T}\right) H_{\mathrm{loc}}(P_R)\varpi^{P_R}_{f,\delta}(\eta_T) \\ &\longrightarrow 0.\end{aligned}$$

This proves $\mathrm{Reg}_{\mathrm{pol}}(T) = o(T)$.

Finally choose $\alpha_T = T^{-2}$. All appearances of $1/\alpha_T$ are inside logarithms, so the preceding limits remain valid. On the failure event use $\mathrm{Reg}_{\mathrm{pol}}(T) \le T$. If $B_T = o(T)$ denotes the good event bound, then

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_{\mathrm{pol}}(T)] &\le B_T + T\alpha_T \\ &= o(T) + T^{-1} \\ &= o(T).\end{aligned}$$

It remains to derive the explicit fixed-face rate. Under the two additional conditions in Theorem 9, $\widehat{P}_T$ and P_R lie in $\mathcal{K}_{\mathrm{face}}(P_R)$ row by row. Hence Lemma 15 gives

$$|\sigma^f_\delta(\widehat{P}_{T,s,a}, u) - \sigma^f_\delta(P_{R,s,a}, u)| \le \mathrm{sp}(u)K_{\mathrm{face}}(P_R)\eta_T.$$

Applying the proof of Lemma 16 with this bound replaces $\varpi_{f,\delta}^{P_R}(\eta_T)$ by $K_{\mathrm{face}}(P_R)\eta_T$ in the regret display.

Lemma 17 gives

$$n_T(P,\alpha) = \widetilde{\Omega}\Big(\frac{T^{2/3}}{D_0(P)d_R A}\Big). \tag{50}$$

The multinomial radius therefore satisfies

$$\eta_T = \widetilde{O}\Big(\sqrt{D_0(P)d_R A[\mathcal{S} + \log(SAT/\alpha)]}\, T^{-1/3}\Big). \tag{51}$$

Substitute equations (50) and (51), together with $m_T = \lceil T^{2/3} \rceil$, into Theorem 9:

$$\begin{aligned}\mathrm{Reg}_{\mathrm{pol}}(T) &\le m_T + 2TH_{\mathrm{loc}}(P_R)K_{\mathrm{face}}(P_R)\eta_T \\ &= \widetilde{O}\Big(\Big[1 + H_{\mathrm{loc}}(P_R)K_{\mathrm{face}}(P_R)\sqrt{D_0(P)d_R A[\mathcal{S} + \log(SAT/\alpha)]}\Big]\, T^{2/3}\Big).\end{aligned}$$

The global navigation cost through transient states and the samples required to observe every positive recurrent edge affect only the hidden finite-horizon threshold, not the exponent of T. □

D.11 A TWO-POINT TOOL FOR THE LOWER BOUNDS

For probability laws P and Q, we use Pinsker's inequality

$$\|\mathsf{P} - \mathsf{Q}\|_{\mathrm{TV}} \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(\mathsf{P}\|\mathsf{Q})}. \tag{52}$$

Theorem 10. *Suppose the rewards agree under P_0, P_1, every stationary policy used by $\mathfrak{A}$ has a scalar robust gain, and $\Delta_{\mathrm{dec}}(P_0, P_1) > 0$. Then*

$$\max_{i\in\{0,1\}} \mathbb{E}_i^{\mathfrak{A}}\, \mathrm{Reg}_{\mathrm{pol}}^{(i)}(T) \ge \frac{T\Delta_{\mathrm{dec}}(P_0,P_1)}{2}\left(1 - \sqrt{\frac{\mathcal{I}_T}{2}}\right)_+ .$$

In particular, if $\mathcal{I}_T \le 1/2$, the right-hand side is at least $T\Delta_{\mathrm{dec}}(P_0,P_1)/4$.

Proof. Include the internal random seed of $\mathfrak{A}$ in its history. Let $H_T^{(i)}$ denote the resulting history under P_i. Take a maximal coupling of $H_T^{(0)}$ and $H_T^{(1)}$. By the defining property of a maximal coupling,

$$\mathbb{P}\Big(H_T^{(0)} = H_T^{(1)}\Big) = 1 - \Big\|\mathbb{P}_0^{\mathfrak{A},T} - \mathbb{P}_1^{\mathfrak{A},T}\Big\|_{\mathrm{TV}}. \tag{53}$$

Write $\mathcal{E} := \{H_T^{(0)} = H_T^{(1)}\}$. On $\mathcal{E}$, the algorithm has received the same observations and uses the same internal random seed in both models. Hence it selects the same stationary policy at every time in both models. Denote the policy used at time t by π_t; an episodic policy is simply repeated for each time in its episode.

The robust optimal gain dominates every robust policy gain, so $\ell_i(\pi) \ge 0$. On $\mathcal{E}$, the definition of Δ_{dec} gives, for every $t < T$,

$$\ell_0(\pi_t) + \ell_1(\pi_t) \ge \Delta_{\mathrm{dec}}(P_0,P_1). \tag{54}$$

Summing inequality (54) over $t = 0, \dots, T-1$ yields

$$\mathrm{Reg}_{\mathrm{pol}}^{(0)}(T) + \mathrm{Reg}_{\mathrm{pol}}^{(1)}(T) \ge T\Delta_{\mathrm{dec}}(P_0,P_1) \qquad \text{on } \mathcal{E}. \tag{55}$$

Both regrets are nonnegative on $\mathcal{E}^c$. Taking expectation under the coupling, whose marginals are the two interaction laws, gives

$$\begin{aligned}&\mathbb{E}_0^{\mathfrak{A}}\, \mathrm{Reg}_{\mathrm{pol}}^{(0)}(T) + \mathbb{E}_1^{\mathfrak{A}}\, \mathrm{Reg}_{\mathrm{pol}}^{(1)}(T) &(56)\\ &\ge T\Delta_{\mathrm{dec}}(P_0,P_1)\Big(1 - \Big\|\mathbb{P}_0^{\mathfrak{A},T} - \mathbb{P}_1^{\mathfrak{A},T}\Big\|_{\mathrm{TV}}\Big). &(57)\end{aligned}$$

At least one of the two expectations is at least half of their sum. Inequality (52) gives

$$\Big\|\mathbb{P}_0^{\mathfrak{A},T} - \mathbb{P}_1^{\mathfrak{A},T}\Big\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\mathrm{KL}\Big(\mathbb{P}_0^{\mathfrak{A},T}\Big\|\mathbb{P}_1^{\mathfrak{A},T}\Big)}.$$

Combining inequality (56) with inequality (52) and taking the positive part proves the first claim. If the KL divergence is at most $1/2$, the upper bound is at most $1/2$, which gives the second claim. □

Theorem 11. *Suppose the rewards agree under P_0, P_1, both robust optimal gains are scalar, and $\Delta^{\mathrm{st}}_{\mathrm{dec}}(P_0, P_1) > 0$. Then every policy-valued learner satisfies*

$$\max_{i\in\{0,1\}} \mathbb{E}_i^{\mathfrak{A}} \operatorname{Reg}^{(i)}_{\mathrm{pol}}(T) \geq \frac{T\Delta^{\mathrm{st}}_{\mathrm{dec}}(P_0, P_1)}{2} \left(1 - \sqrt{\frac{\mathcal{I}_T}{2}}\right)_+ .$$

In particular, if $\mathcal{I}_T \leq 1/2$, the right-hand side is at least $T\Delta^{\mathrm{st}}_{\mathrm{dec}}(P_0, P_1)/4$.

Proof. Use the maximal coupling in the proof of Theorem 10. On the event that the two complete histories agree, both the announced policy π_t and the current state S_t agree in the two models. Therefore, at every time,

$$\ell_0(\pi_t, S_t) + \ell_1(\pi_t, S_t) \geq \Delta^{\mathrm{st}}_{\mathrm{dec}}(P_0, P_1).$$

Summing this statewise inequality and applying the same maximal-coupling and Pinsker bounds proves the result. No scalarity assumption on the gains of suboptimal policies is used. □

Lemma 19. *The map $(q, u) \mapsto \sigma^f_\delta(q, u)$ is continuous on $\Delta(\mathcal{S}) \times \mathbb{R}^{\mathcal{S}}$.*

Proof. Continuity in u is uniform by the row Lipschitz bound in equation (15), so it suffices to fix u and take $q_n \to q$ in $\Delta(\mathcal{S})$.

For the lower bound, let p_n attain $\sigma^f_\delta(q_n, u)$ and pass to a subsequence realizing $\liminf_n \sigma^f_\delta(q_n, u)$ along which $p_n \to p$. Joint lower semicontinuity of $(p, q) \mapsto D_f(p\|q)$ gives $D_f(p\|q) \leq \liminf_n D_f(p_n\|q_n) \leq \delta$, so p is feasible at q and $\liminf_n \sigma^f_\delta(q_n, u) = p^\top u \geq \sigma^f_\delta(q, u)$.

For the upper bound, let p^* attain $\sigma^f_\delta(q, u)$ and fix $\tau \in (0, 1)$. Define

$$p^\tau := (1-\tau)p^* + \tau q, \qquad p^\tau_n := (1-\tau)p^* + \tau q_n.$$

Convexity gives $D_f(p^\tau\|q) \leq (1-\tau)\delta < \delta$. Coordinatewise continuity of the closed perspective yields $D_f(p^\tau_n\|q_n) \to D_f(p^\tau\|q)$: ordinary continuity applies where $q_i > 0$; if $q_i = 0 < p^*_i$, feasibility forces finite recession slope and gives the singular limit; and if $q_i = p^*_i = 0$, the contribution is $q_{n,i} f(\tau) \to 0$. Thus p^τ_n is feasible at q_n for all large n, and

$$\limsup_n \sigma^f_\delta(q_n, u) \leq \lim_n (p^\tau_n)^\top u = (p^\tau)^\top u.$$

Letting $\tau \downarrow 0$ proves the upper bound. □

D.12 Minimax refinements

Proposition 7. *Let $G_\wedge \in \mathcal{M}$ and $G_\wedge \subseteq H \subseteq G$. For $\pi \in \Pi_{\mathrm{d}}$, let $R \in \operatorname{Rec}(G_\pi)$. There is a $\bar\pi \in \Pi_{\mathrm{d}}$ such that*

1. *$\bar\pi = \pi$ on R and $G_{\bar\pi}$ is unichain with recurrent class R;*
2. *every element of $\operatorname{Rec}(H_{\bar\pi})$ is contained in R, for every H satisfying $G_\wedge \subseteq H \subseteq G$.*

If $f \in \mathcal{F}_{\mathrm{ne}}$, $\mathcal{G}(Q) \subseteq G$, $G_\wedge \subseteq \mathcal{G}_\wedge(Q)$, and $\pi \in \operatorname{Gr}_Q(h)$ for $(g, h) \in \operatorname{Eig}(\mathcal{L}^{\mathrm{rob}}_Q)$, then

$$g^{\bar\pi}_{\mathrm{rob},Q}(s) = g, \qquad s \in \mathcal{S}.$$

Thus deleting edges may split R, but it cannot place stationary mass where the completion changed a greedy action.

Proof. We first prove the graph statement by showing that every state can be routed to R in $G_\wedge$, and then prove the gain statement. Since R is a recurrent class of G_π, it is nonempty and

$$G(s, \pi(s)) \subseteq R, \qquad s \in R.$$

The inclusion $G_\wedge \subseteq G$ therefore gives

$$G_\wedge(s, \pi(s)) \subseteq R, \qquad s \in R.$$

Thus R is closable in $G_\wedge$.

Let X be the set of states from which no directed path in $G_\wedge$ reaches R, where the action may be chosen separately at each step. Suppose X were nonempty. It is disjoint from R by definition. We claim that X is closed in $G_\wedge$. Indeed, fix $x \in X$ and $a \in \mathcal{A}$. If there were $y \in G_\wedge(x,a) \setminus X$, then y would have a path to R. Prepending the edge (x,a,y) would give a path from x to R, contrary to $x \in X$. Hence

$$G_\wedge(x,a) \subseteq X \qquad (x \in X, a \in \mathcal{A}),$$

so X is closed. We have obtained a nonempty closed set X and a disjoint nonempty closable set R, contradicting $G_\wedge \in \mathcal{M}$. Therefore $X = \varnothing$.

For every state s, let $d(s)$ be the length of a shortest $G_\wedge$-path from s to R. Thus $d(s) = 0$ exactly on R. For each $s \notin R$, choose an action a_s and a state $y_s \in G_\wedge(s, a_s)$ satisfying

$$d(y_s) = d(s) - 1.$$

Define

$$\bar{\pi}(s) := \begin{cases} \pi(s), & s \in R, \\ a_s, & s \notin R. \end{cases}$$

We next locate every recurrent class. Fix any graph H with $G_\wedge \subseteq H \subseteq G$. Because R is closed under G_π and $\bar{\pi} = \pi$ on R,

$$H(s, \bar{\pi}(s)) \subseteq G(s, \pi(s)) \subseteq R, \qquad s \in R.$$

Thus R is closed under $H_{\bar{\pi}}$.

Suppose K were a recurrent class of $H_{\bar{\pi}}$ disjoint from R. Choose $s \in K$ minimizing $d(s)$. Since recurrent classes are closed,

$$H(s, \bar{\pi}(s)) \subseteq K.$$

On the other hand,

$$y_s \in G_\wedge(s, \bar{\pi}(s)) \subseteq H(s, \bar{\pi}(s)),$$

so $y_s \in K$. This contradicts $d(y_s) = d(s) - 1$ and the minimality of $d(s)$ on K. Hence every recurrent class of $H_{\bar{\pi}}$ is contained in R.

Taking $H = G$, the original policy graph G_π is irreducible on its recurrent class R, and the completion has not changed any action on R. Hence R remains a recurrent class of $G_{\bar{\pi}}$. The preceding paragraph rules out every other recurrent class. Therefore $G_{\bar{\pi}}$ is unichain with recurrent class R.

It remains to prove robust gain preservation. Assume now that $f \in \mathcal{F}_{\mathrm{ne}}$, $\mathcal{G}(Q) \subseteq G$, $G_\wedge \subseteq \mathcal{G}_\wedge(Q)$, and

$$g\mathbf{1} + h = \mathcal{L}_Q^{\mathrm{rob}} h.$$

Let π be greedy for h. Since $f \in \mathcal{F}_{\mathrm{ne}}$, no ambiguity row can create a nonnominal edge. By the definition of the robust core, every $\widetilde{Q} \in \mathcal{U}^Q$ therefore satisfies

$$G_\wedge \subseteq \mathcal{G}(\widetilde{Q}) \subseteq G.$$

By the recurrent-class conclusion above, every recurrent class K of $\widetilde{Q}_{\bar{\pi}}$ is contained in R. On R the completion equals the greedy policy, so for every $s \in K$,

$$\begin{aligned} g + h(s) &= r(s, \bar{\pi}(s)) + \sigma_\delta^f(Q_{s,\bar{\pi}(s)}, h) \\ &\le r(s, \bar{\pi}(s)) + \widetilde{Q}_{s,\bar{\pi}(s)}^\top h. \end{aligned}$$

Let μ_K be the stationary distribution on K. Multiply the last inequality by $\mu_K(s)$ and sum over $s \in K$. Stationarity gives

$$\sum_{s \in K} \mu_K(s) \widetilde{Q}_{s,\bar{\pi}(s)}^\top h = \sum_{s \in K} \mu_K(s) h(s).$$

The two bias averages cancel, and hence

$$g \le \sum_{s \in K} \mu_K(s) r(s, \bar{\pi}(s)). \tag{58}$$

Every trajectory of a finite Markov chain eventually enters one of its recurrent classes with probability one, and its transient contribution to a Cesàro reward average is zero. Inequality (58) for every recurrent class therefore yields

$$g^{\bar{\pi}}_{\widetilde{Q}}(x) \geq g, \qquad x \in \mathcal{S}.$$

Taking the infimum over $\widetilde{Q} \in \mathcal{U}^Q$ gives

$$g^{\bar{\pi}}_{\mathrm{rob},Q}(x) \geq g.$$

Inequality (19) in the proof of Theorem 3 uses only the existence of the eigenpair (g, h). Applied to the kernel assembled from the rowwise minimizers of h, it gives $g^{\pi}_{\mathrm{rob},Q} \leq g$ for every π, so no policy can have robust gain strictly larger than g. Consequently

$$g^{\bar{\pi}}_{\mathrm{rob},Q}(x) = g \qquad (x \in \mathcal{S}),$$

as claimed. □

Proposition 8. *Let $f \in \mathcal{F}_{\mathrm{ne}}$, let $\delta < \delta^{\mathrm{bell}}_f(Q)$, and let $\pi \in \Pi_{\mathrm{d}}$. For every $R \in \mathrm{Rec}(\mathcal{G}(Q)_\pi)$ there is a deterministic policy $\bar{\pi}$ such that*

1. *$\bar{\pi} = \pi$ on R and $\mathcal{G}(Q)_{\bar{\pi}}$ is unichain with recurrent class R;*
2. *every recurrent class of $\widetilde{Q}_{\bar{\pi}}$ is contained in R, for every $\widetilde{Q} \in \mathcal{U}^Q(\delta)$.*

If $\pi \in \mathrm{Gr}_Q(h)$ for $(g, h) \in \mathrm{Eig}(\mathcal{L}^{\mathrm{rob}}_Q)$, then

$$g^{\bar{\pi}}_{\mathrm{rob},Q}(x) = g, \qquad x \in \mathcal{S}.$$

More generally, let $\mathcal{C}(s, a)$ be nonempty row sets that all have the same support graph G, and suppose

$$\forall I, J \neq \varnothing, I \cap J = \varnothing,\ J \text{ closable in } G, \quad \exists (s, a) \in I \times \mathcal{A} : \quad \inf_{q \in \mathcal{C}(s,a)} q(I^c) > q_\star. \tag{59}$$

Then one completion, obtained using only the certified inequalities in (59)*, has conclusions 1–2 simultaneously for every center $Q \in \prod_{s,a} \mathcal{C}(s, a)$ and every ambiguity kernel around that center, whenever R is a recurrent class of G_π.*

Proof. Define $X_0 := \mathcal{S} \setminus R$. While X_{j-1} is nonempty, the set R is nominally closable and disjoint from X_{j-1}. Since $\delta < \delta^{\mathrm{bell}}_f(Q)$, the pair (X_{j-1}, R) cannot be a Bellman obstruction. Hence there are $s_j \in X_{j-1}$ and $a_j \in \mathcal{A}$ such that

$$\theta_f\big(Q(X^c_{j-1} \mid s_j, a_j)\big) > \delta, \qquad Q(X^c_{j-1} \mid s_j, a_j) > q_\star. \tag{60}$$

Set $\bar{\pi}(s_j) := a_j$ and $X_j := X_{j-1} \setminus \{s_j\}$, and keep $\bar{\pi} = \pi$ on R. This deletes one state per step and therefore defines $\bar{\pi}$ on all of $\mathcal{S}$.

No row in the f-ball centered at Q_{s_j,a_j} can be supported inside X_{j-1}: doing so would delete the nominal mass in X^c_{j-1}, whose exact cost in (60) exceeds δ. Suppose nevertheless that some ambiguity kernel had a recurrent class K disjoint from R. Choose the earliest removed state $s_j \in K$. Then $K \subseteq X_{j-1}$, whereas closedness of K would force the row at (s_j, a_j) to be supported inside X_{j-1}, a contradiction. Because no creation is possible and π closes R in $\mathcal{G}(Q)$, every ambiguity row selected on R also stays in R. Consequently every ambiguity recurrent class is contained in R. For the center Q, the irreducible class R remains recurrent and no other recurrent class exists, proving conclusions 1–2.

If π is greedy, repeat the stationary-drift argument in the last part of Proposition 7. Every adversarial stationary distribution is supported on R, where $\bar{\pi} = \pi$, so greedy equality gives $g^{\bar{\pi}}_{\mathrm{rob},Q}(x) \geq g$ for every x. The rowwise-minimizer kernel gives the reverse inequality for every policy, and hence equality.

Under (59), use at each elimination step the lexicographically first certified pair (s_j, a_j). Every row in every $\mathcal{C}(s_j, a_j)$ has outgoing mass greater than $q_\star$ and therefore deletion cost greater than δ. The preceding earliest-removal proof is then simultaneous over all centers and all their ambiguity kernels. Since all centers have graph G, the same R is closable throughout and the construction is well defined. □

Algorithm 2 Optimistic completion (OC)

Input: T, α, δ; $m_T^{\mathrm{OC}} := \lceil \sqrt{T} \rceil$
Play $A_t \sim \pi_0(\cdot \mid S_t)$ for $t < m_T^{\mathrm{OC}}$
for each doubling episode k with $t_k \geq m_T^{\mathrm{OC}}$ **do**
  Form B_k, the sets $\mathcal{C}^{\circ}_{k,B_k}(s,a)$, and $\mathsf{Ready}_k(B_k)$ as in equation (22)
  if $\mathsf{Ready}_k(B_k)$ fails **then**
    Play π_0 until some $N(s,a) - N_k(s,a)$ reaches $N_k(s,a) \vee 1$; **continue**
  Choose an optimistic center $Q_k \in \arg\max\{g^{\mathrm{rob}}(Q) : Q_{s,a} \in \mathcal{C}^{\circ}_{k,B_k}(s,a),\ \mathrm{supp}(Q_{s,a}) = \widehat{G}_k(s,a) \cap B_k\ \forall (s,a) \in B_k \times \mathcal{A}\}$
  (the support constraint is imposed whenever $\mathsf{CoreReady}_k(B_k)$ holds through its one-sided branch)
  Set $g_k := g^{\mathrm{rob}}(Q_k)$, take $h_k \in \mathcal{H}(Q_k)$ as a $\gamma \uparrow 1$ accumulation point of $v_{Q_k,\gamma} - v_{Q_k,\gamma}(S_{t_k})\mathbf{1}$, and take $\pi_k \in \mathrm{Gr}_{Q_k}(h_k)$
  Select a recurrent class R_k of $(\widehat{G}_k|_{B_k})_{\pi_k}$ by a fixed deterministic rule
  Let $G_k^{\wedge} := H_k^-|_{B_k}$ if $\mathsf{CoreReady}_k(B_k)$ holds through its two-sided branch, and $G_k^{\wedge} := \widehat{G}_k|_{B_k}$ if it holds through the one-sided branch
  On B_k, let $\bar{\pi}_k$ be the completion in Proposition 7 with $G_{\wedge} = G_k^{\wedge}$, $H = H_k^-|_{B_k}$, $G = \widehat{G}_k|_{B_k}$, and recurrent class R_k; set $\bar{\pi}_k = \pi_0$ on $\mathcal{S} \setminus B_k$
  Play $\bar{\pi}_k$ until some $N(s,a) - N_k(s,a)$ reaches $N_k(s,a) \vee 1$

Algorithm 3 Optimistic Bellman completion (BC)

Input: T, α, δ; $m_T^{\mathrm{BC}} := \lceil \sqrt{T} \rceil$
Play $A_t \sim \pi_0(\cdot \mid S_t)$ for $t < m_T^{\mathrm{BC}}$
for each doubling episode k with $t_k \geq m_T^{\mathrm{BC}}$ **do**
  Form B_k, $\widehat{G}_k$, and the sets $\mathcal{C}^{\circ}_{k,B_k}(s,a)$
  if $\mathsf{BellReady}_k(B_k)$ in (23) fails **then**
    Play π_0 until some $N(s,a) - N_k(s,a)$ reaches $N_k(s,a) \vee 1$; **continue**
  Choose $Q_k \in \arg\max\{g^{\mathrm{rob}}(Q) : Q_{s,a} \in \mathcal{C}^{\circ}_{k,B_k}(s,a)\ \ \forall (s,a) \in B_k \times \mathcal{A}\}$
  Take an eigenpair (g_k, h_k) and a greedy selector $\pi_k \in \mathrm{Gr}_{Q_k}(h_k)$ as in Algorithm 2
  Select $R_k \in \mathrm{Rec}((\widehat{G}_k|_{B_k})_{\pi_k})$ by a fixed rule; initialize $X \leftarrow B_k \setminus R_k$ and $\bar{\pi}_k \leftarrow \pi_k$ on R_k
  while $X \neq \varnothing$ **do**
    Choose the first $(s,a) \in X \times \mathcal{A}$ satisfying $\inf_{q \in \mathcal{C}^{\circ}_{k,B_k}(s,a)} q(B_k \setminus X) > q_\star$
    Set $\bar{\pi}_k(s) \leftarrow a$ and $X \leftarrow X \setminus \{s\}$
  Set $\bar{\pi}_k = \pi_0$ on $\mathcal{S} \setminus B_k$
  Play $\bar{\pi}_k$ until some $N(s,a) - N_k(s,a)$ reaches $N_k(s,a) \vee 1$

Theorem 12. *Let $f \in \mathcal{F}_{\mathrm{ne}}$, $\mathcal{G}(P) \in \mathcal{M}$, $\delta < \delta_f^*(P)$, and $D_0(P) + D_{\mathrm{cmp}}(P_R) + H_{\mathrm{loc}}(P_R) < \infty$. If $n_T^{\mathrm{OC}}(P,\alpha) \geq n_{\mathrm{ne}}(P,\alpha)$, then Algorithm 2 satisfies*

$$\mathbb{E}\,\mathrm{Reg}_{\mathrm{pol}}(T) \leq m_T^{\mathrm{OC}} + \widetilde{O}\Big(\mathfrak{C}_{\mathrm{ne}}(P_R)\sqrt{SAT\,[S + \log(SAT/\alpha)]}\Big) + \widetilde{O}\big(\mathfrak{C}_{\mathrm{ne}}(P_R) M_{\mathrm{mix}}(P_R) S^2 A\big) + \alpha T,$$

where $\mathfrak{C}_{\mathrm{ne}}(P_R) := H_{\mathrm{loc}}(P_R) K_{\mathrm{face}}(P_R) C_{\mathrm{occ}}(P_R)$. The unrestricted no-creation minimax conclusion is supplied by Theorem 13.

Proof. Write $S := |\mathcal{S}|$, $A := |\mathcal{A}|$,

$$H := H_{\mathrm{loc}}(P_R), \qquad K := K_{\mathrm{face}}(P_R), \qquad C := C_{\mathrm{occ}}(P_R), \qquad M := M_{\mathrm{mix}}(P_R).$$

We prove that $C, M < +\infty$, derive the episodewise occupation inequality, and then sum it.

We first prove that the two occupation constants are finite. The set $\Pi_{\mathrm{cmp}}(P_R)$ is finite because it is a subset of the finite set of deterministic policies on $R(P)$; it is nonempty by Proposition 8. Fix $\pi \in \Pi_{\mathrm{cmp}}(P_R)$ and let R_π be the unique recurrent class of $\mathcal{G}(P_R)_\pi$. The stationary distribution $\mu^\pi_{P_R}$

is strictly positive on every state of R_π. By the definition of $\Pi_{\mathrm{cmp}}(P_R)$, every recurrent class of $\widetilde{Q}_\pi$ is contained in R_π for every $\widetilde{Q} \in \mathcal{U}^{P_R}$. It follows that every stationary distribution of $\widetilde{Q}_\pi$ is supported on R_π. Hence, for every stationary μ and every $s \in R_\pi$,

$$\frac{\mu(s)}{\mu^\pi_{P_R}(s)} \le \frac{1}{\mu^\pi_{P_R}(s)}.$$

Taking a maximum over finitely many policies and recurrent states proves

$$C \le \max_{\pi \in \Pi_{\mathrm{cmp}}(P_R)} \max_{s \in R_\pi} \frac{1}{\mu^\pi_{P_R}(s)} < +\infty.$$

For M, fix $\pi \in \Pi_{\mathrm{cmp}}(P_R)$ and $s \in R(P)$. The centered reward

$$d_s(x) := \mathbf{1}\{x = s\} - \mu^\pi_{P_R}(s)$$

has zero stationary mean on the unique recurrent class. The finite-state Poisson equation

$$v_s - (P_R)_\pi v_s = d_s$$

has a bounded solution on that recurrent class. Extend it to the transient states by the usual first-entry representation. The extension is finite because a finite unichain has finite expected entrance time to its recurrent class from every state. Telescoping the Poisson equation gives, for every x and n,

$$\mathbb{E}^\pi_{P_R}\left[\sum_{t<n} d_s(S_t) \mid S_0 = x\right] = v_s(x) - \mathbb{E}^\pi_{P_R}[v_s(S_n) \mid S_0 = x]. \tag{61}$$

The absolute value is at most $2\|v_s\|_\infty$. Maximizing over the finitely many pairs (π, s) proves $M < +\infty$. Normalize each solution by requiring its stationary mean to be zero. The Cesàro averages of $\mathbb{E}_x[v_s(S_n)]$ then converge to zero. Averaging equation (61) over $n = 1, \dots, N$ and using the definition of M gives $|v_s(x)| \le M$ as $N \to \infty$. Hence every endpoint difference satisfies $|v_s(y) - v_s(x)| \le 2M$.

We next verify the exploration and readiness event. Apply Lemma 17 with m_T^{OC} in place of m_T and $n_T^{\mathrm{OC}}(P, \alpha)$ in place of $n_T(P, \alpha)$. Under the sample threshold in Theorem 12, with probability at least $1 - \alpha$ all recurrent rows have at least $n_{\mathrm{ne}}(P, \alpha)$ observations, every positive edge of P_R has been observed, and all row confidence statements hold. Consequently $B_{k_T} = C(P)$, and the same equality holds at every later episode because the nominal process cannot leave $C(P)$.

Write

$$n_{\mathrm{cert}} := n^\sharp_{\mathrm{cert}}(P, \alpha),$$
$$n_{\mathrm{face}} := \min\{n \ge 1 : 2\beta_{\bar\alpha}(n) \le \bar\epsilon_{\mathrm{face}}(P_R)\}.$$

If $q_\star = 0$ and $C(P) = \mathcal{S}$, support recovery gives $B_{k_T} = \mathcal{S}$ and $\widehat{G}_{k_T} = \mathcal{G}(P)$. The latter graph is weakly communicating, so $\mathsf{CoreReady}_{k_T}(B_{k_T})$ holds by its one-sided branch. In all other cases, $n \ge n_{\mathrm{cert}}$ gives $C(P) \in \mathfrak{B}_{k_T}$ by Proposition 6, and hence the first branch holds. Support recovery also gives $\widehat{G}_{k_T}|_{B_{k_T}} = \mathcal{G}(P_R)$ and makes $P_R \in \prod_{s,a} \mathcal{C}^\circ_{k_T, B_{k_T}}(s, a)$, so the product is nonempty and $H^-_{k_T}|_{B_{k_T}} \subseteq \widehat{G}_{k_T}|_{B_{k_T}}$. Finally, for every row $q \in \mathcal{C}^\circ_{k_T, B_{k_T}}(s, a)$, the confidence event and triangle inequality give

$$\|q - P_{R,s,a}\|_1 \le 2\beta_{\bar\alpha}(N_{k_T}(s, a)) \le \bar\epsilon_{\mathrm{face}}(P_R).$$

Thus the entire product lies in $\mathcal{K}_{\mathrm{face}}(P_R) \subseteq \mathcal{K}(P_R)$, on which $\delta < \delta^{\mathrm{bell}}_f(Q)$. This proves $\mathsf{Ready}_{k_T}(B_{k_T})$. Intersected confidence sets shrink with k, and the recovered support and certificate persist, so every later optimistic center remains on the same nominal support faces. The algorithm therefore never invokes its fallback on this event.

Inverting the thresholds gives

$$n_{\mathrm{cert}} = \begin{cases} \widetilde{O}(p_{\min,R}(P)^{-1}), & q_\star = 0,\ C(P) = \mathcal{S}, \\ \widetilde{O}(\epsilon_{\mathrm{top}}(P)^{-2}), & \text{otherwise}, \end{cases}$$
$$n_{\mathrm{face}} = \widetilde{O}(\bar\epsilon_{\mathrm{face}}(P_R)^{-2}),$$
$$n_{\mathrm{supp}}(P, \alpha) = \widetilde{O}(p_{\min,R}(P)^{-1}),$$

so $n_{\mathrm{ne}}(P,\alpha) = \widetilde{O}(\Upsilon_{\mathrm{ne}}(P))$. The factor two in n_{face} is essential because an arbitrary confidence-set row and the true row may lie on opposite sides of the empirical center.

We next localize the analysis at episode starts. Let E_{init} be the event that the exploration phase ends in $C(P)$, observes every positive edge of P_R, and supplies at least $n_{\mathrm{ne}}(P,\alpha)$ samples per recurrent row. For each post-exploration episode, let E_k be E_{init} intersected with validity of all row confidence statements through t_k. Then $E_k \in \mathcal{F}_{t_k}$, the preceding readiness conclusions hold on E_k, and the joint exploration and confidence guarantee gives $\mathbb{P}(\exists k : E_k^c) \le \alpha$, where only episodes starting before T are included. Define the actual episode regret

$$\mathcal{R}_k := \sum_{t=t_k}^{t_k+L_k(T)-1} [g^{\mathrm{rob}}(P) - g^{\pi_t}_{\mathrm{rob},P}(S_t)].$$

Fix an episode-start history in E_k. All subsequent conditional expectations use the original nominal law given $\mathcal{F}_{t_k}$; no future confidence event is conditioned on. Let Q_k be its optimistic center, (g_k, h_k) an additive eigenpair, π_k a greedy selector, and $\bar{\pi}_k$ the completion in Proposition 7. Let R_k be the recurrent class selected from the observed support graph on B, which equals the nominal support graph on the event under consideration. By construction,

$$\bar{\pi}_k(s) = \pi_k(s), \qquad s \in R_k.$$

The recovered support and certified-core event makes the graph sandwich in Proposition 7 valid simultaneously for Q_k, P, and their ambiguity kernels.

Normalize h_k so that $\min_s h_k(s) = 0$. Then

$$\mathrm{sp}(h_k) \le H.$$

For $s \in R_k$, greedy equality gives

$$g_k + h_k(s) = r(s, \bar{\pi}_k(s)) + \sigma^f_\delta((Q_k)_{s,\bar{\pi}_k(s)}, h_k).$$

The confidence event and the triangle inequality give

$$\|(Q_k)_{s,a} - P_{s,a}\|_1 \le 2\beta_{\bar{\alpha}}(N_k(s,a)).$$

The fixed-face modulus bound from Lemma 15, together with homogeneity in h_k, therefore implies

$$\begin{aligned}\sigma^f_\delta((Q_k)_{s,a}, h_k) - \sigma^f_\delta(P_{s,a}, h_k) &\le \mathrm{sp}(h_k) K \|(Q_k)_{s,a} - P_{s,a}\|_1 \\ &\le 2HK\beta_{\bar{\alpha}}(N_k(s,a)).\end{aligned}$$

Consequently, for $s \in R_k$ and $a = \bar{\pi}_k(s)$,

$$\begin{aligned}g_k + h_k(s) \le r(s,a) + \sigma^f_\delta(P_{s,a}, h_k) \\ + 2HK\beta_{\bar{\alpha}}(N_k(s,a)).\end{aligned} \tag{62}$$

Fix an initial state x for the episode and an arbitrary $\varepsilon > 0$. Choose $\widetilde{Q}_{k,\varepsilon} \in \mathcal{U}^{P_R}$ whose average reward under $\bar{\pi}_k$ from x is within ε of the infimum defining $g^{\bar{\pi}_k}_{\mathrm{rob},P}(x)$. Let

$$\mu_{k,\varepsilon} := \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathcal{L}^{\bar{\pi}_k}_{\widetilde{Q}_{k,\varepsilon}}(S_t \mid S_0 = x)$$

be the Cesàro limiting state distribution. This limit exists for every finite Markov chain, is stationary, and gives the average reward:

$$\begin{aligned}\mu^\top_{k,\varepsilon}(\widetilde{Q}_{k,\varepsilon})_{\bar{\pi}_k} &= \mu^\top_{k,\varepsilon}, \\ g^{\bar{\pi}_k}_{\widetilde{Q}_{k,\varepsilon}}(x) &= \sum_s \mu_{k,\varepsilon}(s) r(s, \bar{\pi}_k(s)).\end{aligned}$$

Proposition 7 shows that this stationary distribution is supported on R_k. Also, for every s,

$$\sigma^f_\delta(P_{s,\bar{\pi}_k(s)}, h_k) \le (\widetilde{Q}_{k,\varepsilon})^\top_{s,\bar{\pi}_k(s)} h_k.$$

Multiply inequality (62) by $\mu_{k,\varepsilon}(s)$ and sum. The support property permits use of inequality (62) at every state with positive weight, and stationarity cancels the bias terms. We obtain

$$g_k \le g^{\bar\pi_k}_{\widetilde{Q}_{k,\varepsilon}}(x) + 2HK\sum_s \mu_{k,\varepsilon}(s)\beta_{\bar\alpha}(N_k(s,\bar\pi_k(s))).$$

The construction applies to every $x \in R(P)$, and its right-hand side is independent of x. Letting $\varepsilon \downarrow 0$ and using the definition of C therefore yields, for every $x \in R(P)$,

$$g_k - g^{\bar\pi_k}_{\mathrm{rob},P_R}(x) \le 2HKC\sum_s \mu^{\bar\pi_k}_{P_R}(s)\beta_{\bar\alpha}(N_k(s,\bar\pi_k(s))). \tag{63}$$

Because P_R belongs to the support-restricted confidence product, optimism gives $g^{\mathrm{rob}}(P_R) \le g_k$. Hence the left side of inequality (63) upper-bounds the episode's robust policy gap at every $x \in R(P)$. By Lemma 18, $g^{\mathrm{rob}}(P_R) = g^{\mathrm{rob}}(P)$ and the restricted and unrestricted robust policy gains agree at every state of $C(P)$. Consequently the same state-free right-hand side bounds every summand $g^{\mathrm{rob}}(P) - g^{\bar\pi_k}_{\mathrm{rob},P}(S_t)$ during episode k.

We now replace nominal stationary masses by visit counts. Let $n_k(s,a)$ be the number of visits to (s,a) during episode k. For an episode-start history in E_k, the policy $\bar\pi_k$ is fixed and deterministic given $\mathcal{F}_{t_k}$, and $\bar\pi_k \in \Pi_{\mathrm{cmp}}(P_R)$ by Proposition 7. The episode length is a stopping time, so the deviation constant M must be applied through optional stopping rather than through a bound uniform in a deterministic number of steps.

Let v_s solve the Poisson equation $v_s - (P_R)_{\bar\pi_k}v_s = \mathbf{1}\{\cdot = s\} - \mu^{\bar\pi_k}_{P_R}(s)$ constructed above and normalized to have zero stationary mean. Then

$$Z_n := \sum_{t<n}\big(\mathbf{1}\{S_t = s\} - \mu^{\bar\pi_k}_{P_R}(s)\big) + v_s(S_n) - v_s(S_0)$$

is a martingale with bounded increments under the fixed episode policy (time is reset to zero at t_k). Applying optional stopping at the stopping time $\tau := \tau_k \wedge (T - t_k)$, whose value is $L_k(T)$ and whose expectation is finite, gives $\mathbb{E}[Z_\tau \mid \mathcal{F}_{t_k}] = 0$ and hence

$$\mathbb{E}[L_k(T) \mid \mathcal{F}_{t_k}]\,\mu^{\bar\pi_k}_{P_R}(s)\mathbf{1}\{a = \bar\pi_k(s)\} \tag{64}$$

$$\le \mathbb{E}[n_k(s,a) \mid \mathcal{F}_{t_k}] + 2M, \tag{65}$$

because $|v_s(S_\tau) - v_s(S_0)| \le 2M$. Sum the uniform bound (63) over the realized states in episode k and take conditional expectations. Its state-free right-hand side is $\mathcal{F}_{t_k}$-measurable, so the resulting factor is $\mathbb{E}[L_k(T) \mid \mathcal{F}_{t_k}]$, and inequality (64) applies term by term. Since $0 \le \beta_{\bar\alpha} \le 2$ and $E_k \in \mathcal{F}_{t_k}$, multiplying by $\mathbf{1}_{E_k}$ and taking unconditional expectations gives

$$\mathbb{E}[\mathbf{1}_{E_k}\mathcal{R}_k] \le 2HKC\sum_{s,a}\mathbb{E}[\mathbf{1}_{E_k}n_k(s,a)\beta_{\bar\alpha}(N_k(s,a))] + 8HKCMS\mathbb{P}(E_k). \tag{66}$$

This bound charges the entire episode whenever its starting history is valid, even if a confidence statement fails later in that episode.

We next bound the doubling sum. Let

$$L_T := \sqrt{2\log(c_{\bar\alpha}T(T+1))}.$$

For every $n \le T$,

$$\beta_{\bar\alpha}(n) \le \min\left\{2, \frac{L_T}{\sqrt{n \vee 1}}\right\}.$$

Under the doubling stopping rule,

$$0 \le n_k(s,a) \le N_k(s,a) \vee 1.$$

For $N \ge 1$ and $0 \le n \le N$,

$$\sqrt{N+n} - \sqrt{N} = \frac{n}{\sqrt{N+n} + \sqrt{N}} \tag{67}$$

$$\ge \frac{n}{(1+\sqrt{2})\sqrt{N}}. \tag{68}$$

The first visit to a row contributes at most two. Applying inequality (67) to every later episode and telescoping gives, pathwise,

$$\sum_k n_k(s,a)\beta_{\bar{\alpha}}(N_k(s,a)) \le 2 + (1+\sqrt{2})L_T\sqrt{N_T(s,a)}.$$

Summing over rows and using $\sum_{s,a}\sqrt{N_T(s,a)} \le \sqrt{SA\sum_{s,a}N_T(s,a)}$,

$$\sum_{s,a}\sum_k n_k(s,a)\beta_{\bar{\alpha}}(N_k(s,a)) \tag{69}$$

$$\le 2SA + (1+\sqrt{2})L_T\sum_{s,a}\sqrt{N_T(s,a)} \tag{70}$$

$$\le 2SA + (1+\sqrt{2})L_T\sqrt{SAT}, \tag{71}$$

because $\sum_{s,a}N_T(s,a) = T$.

The number K_T of doubling episodes is at most

$$K_T \le 1 + SA\lceil\log_2 T\rceil. \tag{72}$$

Indeed, after its first visit, a row can trigger at most $\lceil\log_2 T\rceil$ doublings before its count exceeds T. Sum inequality (66) over episodes, padding by zero contributions after T up to the deterministic bound in (72). Since the count terms are nonnegative, their indicators can be dropped. The pathwise count bounds (69)–(72) give

$$\mathbb{E}\left[\sum_k \mathbf{1}_{E_k}\mathcal{R}_k\right] \le 2HKC\left[2SA + (1+\sqrt{2})L_T\sqrt{SAT}\right] + 8HKCMS\left[1 + SA\lceil\log_2 T\rceil\right].$$

Every per-step policy gap lies in $[0,1]$, so $\sum_k \mathbf{1}_{E_k^c}\mathcal{R}_k \le T\mathbf{1}_{\{\exists k: E_k^c\}}$ and its expectation is at most αT. Adding the exploration cost m_T^{OC} bounds $\mathbb{E}\,\mathrm{Reg}_{\mathrm{pol}}(T)$. Absorbing logarithms and fixed lower-order factors into $\widetilde{O}$ proves the bound in Theorem 12.

This proves the stated instance-dependent common-core bound. Uniformity on the full no-creation family, including the support-preserving case with transient nominal states for which the one-sided core certificate can be vacuous, follows from the Bellman-completion theorem proved next. □

Theorem 13. *Let $f \in \mathcal{F}_{\mathrm{ne}}$, $\delta < \delta_f^{\mathrm{bell}}(P)$, and $D_0(P) + D_{\mathrm{cmp}}(P_R) + H_{\mathrm{loc}}(P_R) < \infty$. If $n_T^{\mathrm{OC}}(P,\alpha) \ge n_{\mathrm{BC}}(P_R,\alpha)$, then Algorithm 3 satisfies*

$$\mathbb{E}\,\mathrm{Reg}_{\mathrm{pol}}(T) \le m_T^{\mathrm{BC}} + \widetilde{O}\Big(\mathfrak{C}_{\mathrm{ne}}(P_R)\sqrt{d_R AT[S + \log(SAT/\alpha)]}\Big) + \widetilde{O}\big(\mathfrak{C}_{\mathrm{ne}}(P_R)M_{\mathrm{mix}}(P_R)d_R^2 A\big) + \alpha T, \tag{73}$$

where $\mathfrak{C}_{\mathrm{ne}}(P_R) := H_{\mathrm{loc}}(P_R)K_{\mathrm{face}}(P_R)C_{\mathrm{occ}}(P_R)$. Consequently every nonempty scalar no-creation cell has minimax horizon exponent $1/2$.

Proof. We isolate why the proof of Theorem 12 extends beyond the common-core radius and why nominal communication is unnecessary. The inequality $\delta < \delta_f^{\mathrm{bell}}(P)$ implies $\mathcal{G}(P) \in \mathcal{M}$: otherwise a nominal closed–closable obstruction has all-row deletion cost zero and forces $\delta_f^{\mathrm{bell}}(P) = 0$. Lemma 18 gives $\delta < \delta_f^{\mathrm{bell}}(P_R)$ and identifies the robust optimal and policy gains on $C(P)$ with those of P_R. Define E_{init}, $E_k \in \mathcal{F}_{t_k}$, and $\mathcal{R}_k$ as in the preceding proof, using m_T^{BC} and n_{BC} in place of m_T^{OC} and n_{ne}. The exploration and confidence guarantee again gives $\mathbb{P}(\exists k : E_k^c) \le \alpha$. On E_k, support recovery gives $B_k = C(P)$ and $\widehat{G}_k|_{B_k} = \mathcal{G}(P_R)$. Every restricted confidence row has the same support because

$$2\beta_{\bar{\alpha}}(N_k(s,a)) \le \tfrac{1}{16}\Delta_{\mathrm{bell}}(P_R) \le \tfrac{1}{16}p_{\min,R}(P).$$

For any Bellman obstruction pair (I,J), choose a true row with $P_R(C(P)\backslash I \mid s,a) - q_\star \ge m_{\mathrm{bell}}(P_R)$. For every row q in its restricted confidence set,

$$q(C(P)\setminus I) \ge P_R(C(P)\setminus I \mid s,a) - \tfrac{1}{2}\|q - P_{R,s,a}\|_1 > q_\star.$$

Thus $\mathsf{BellReady}_k(B_k)$ holds. The same radius bound and Lemma 11 place every optimistic center in $\mathcal{K}_{\mathrm{face}}(P_R)$, so its eigenbias has span at most $H_{\mathrm{loc}}(P_R)$.

Fix an episode-start history in E_k. Proposition 8, applied to the confidence product, gives simultaneously

1. $\bar{\pi}_k = \pi_k$ on the selected greedy recurrent class R_k;
2. $(P_R)_{\bar{\pi}_k}$ is unichain with recurrent class R_k;
3. every stationary distribution under every kernel in $\mathcal{U}^{P_R}(\delta)$ is supported on R_k.

In particular $\bar{\pi}_k|_{C(P)} \in \Pi_{\mathrm{cmp}}(P_R)$. These are precisely the properties of the common-core completion used after the readiness part of the proof of Theorem 12. Greedy equality on R_k and the fixed-face modulus give inequality (62). Fix $x \in C(P)$ and $\varepsilon > 0$, and choose $\widetilde{Q}_{k,\varepsilon,x} \in \mathcal{U}^{P_R}$ whose average reward under $\bar{\pi}_k$ from x is at most $g^{\bar{\pi}_k}_{\mathrm{rob},P_R}(x) + \varepsilon$. Let

$$\mu_{k,\varepsilon,x} := \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathcal{L}^{\bar{\pi}_k}_{\widetilde{Q}_{k,\varepsilon,x}}(S_t \mid S_0 = x).$$

For a finite Markov chain this Cesàro limit exists, is stationary, and its reward average equals $g^{\bar{\pi}_k}_{\widetilde{Q}_{k,\varepsilon,x}}(x)$. The third completion property gives $\mathrm{supp}(\mu_{k,\varepsilon,x}) \subseteq R_k$. Multiplying inequality (62) by this distribution, using rowwise feasibility

$$\sigma^f_\delta(P_{R,s,\bar{\pi}_k(s)}, h_k) \le (\widetilde{Q}_{k,\varepsilon,x})^\top_{s,\bar{\pi}_k(s)} h_k,$$

and canceling the bias terms by stationarity gives

$$g_k - g^{\bar{\pi}_k}_{\mathrm{rob},P_R}(x) \le \varepsilon + \sum_s \mu_{k,\varepsilon,x}(s) b_k(s, \bar{\pi}_k(s)).$$

Because $\bar{\pi}_k \in \Pi_{\mathrm{cmp}}(P_R)$, the definition of $C = C_{\mathrm{occ}}(P_R)$ gives

$$\mu_{k,\varepsilon,x}(s) \le C \mu^{\bar{\pi}_k}_{P_R}(s).$$

Equivalently, this uses only a uniform lower bound on the nominal stationary mass over the finite comparison-policy class: the numerator is at most one, while Kac's formula and Lemma 10 give $\mu^{\bar{\pi}_k}_{P_R}(s) \ge (1 + D_{\mathrm{cmp}}(P_R))^{-1}$ on R_k. Apply the confidence-width bound $b_k(s,a) \le 2HK\beta_{\bar{\alpha}}(N_k(s,a))$, use optimism, and send $\varepsilon \downarrow 0$. For every $x \in C(P)$ this yields

$$\begin{aligned} g^{\mathrm{rob}}(P_R) - g^{\bar{\pi}_k}_{\mathrm{rob},P_R}(x) &\le g_k - g^{\bar{\pi}_k}_{\mathrm{rob},P_R}(x) \\ &\le 2HKC \sum_s \mu^{\bar{\pi}_k}_{P_R}(s) \beta_{\bar{\alpha}}(N_k(s, \bar{\pi}_k(s))). \end{aligned} \tag{74}$$

This is the state-uniform episode gap (63). On E_k the nominal process is in $C(P)$ and cannot leave it, so Lemma 18 transfers this bound to every actual regret summand in episode k. The unichain property permits the same Poisson equation and optional-stopping argument under the original nominal law given $\mathcal{F}_{t_k}$. Since E_k is episode-start measurable, (66) holds with $S = d_R$. Only episodes with valid starts contribute; on E_{init} all post-exploration visits and doubling triggers lie in $C(P) \times \mathcal{A}$. Summing and using the pathwise doubling estimates therefore yields

$$\begin{aligned} \mathbb{E}\left[\sum_k \mathbf{1}_{E_k} \mathcal{R}_k\right] &\le \widetilde{O}\Big(HKC\sqrt{d_R AT[S + \log(SAT/\alpha)]}\Big) \\ &\quad + \widetilde{O}\big(HKCMd_R^2 A\big), \end{aligned}$$

where $H = H_{\mathrm{loc}}(P_R)$, $K = K_{\mathrm{face}}(P_R)$, $C = C_{\mathrm{occ}}(P_R)$, and $M = M_{\mathrm{mix}}(P_R)$. The remaining episodes have total regret at most $T\mathbf{1}_{\{\exists k: E_k^c\}}$, whose expectation is at most αT. Adding the exploration cost m_T^{BC} proves (73).

On $\mathfrak{P}_f(B)$, $\mathfrak{P}_f^{\mathrm{all}}(B)$, and $\mathfrak{P}_f^{\mathrm{bell}}(B)$, the constants H, K, C, M are bounded by the same explicit functions of B as in the preceding proof. On the core family, inequality (21) gives $\Delta_{\mathrm{bell}}(P_R)^{-1} \le$

$\Gamma_f(B)$; on the two annular families, $\Delta_{\mathrm{bell}}(P_R) \geq B^{-1}$ holds by definition. Thus $n_{\mathrm{BC}}(P_R, T^{-2}) = \widetilde{O}_{f,\delta}(B^2)$ uniformly over all three cells. The uniform-exploration inversion therefore verifies the burn-in for all sufficiently large T and yields the $\widetilde{O}(\sqrt{T})$ upper bound throughout the scalar no-creation regime. Propositions 9 and 10 supply the matching lower bounds in the core, all, and bell regimes. □

Proposition 9.

$$(\forall f \in \mathcal{F}_{\mathrm{ne}})(\forall \delta < \infty)(\exists B, c_{f,\delta}, T_0 > 0)(\forall T \geq T_0):$$
$$\mathfrak{R}_f(T; B) \;\geq\; c_{f,\delta}\sqrt{T},$$

witnessed by a communicating two-state family with a strict robust core margin. More sharply, for $S := |\mathcal{S}| \geq 3$, $A := |\mathcal{A}| \geq 2$, and $M := \lfloor S/3 \rfloor$, there are $c_{f,\delta}, C_{f,\delta} > 0$ such that

$$B \geq C_{f,\delta} SA, \quad T \geq C_{f,\delta} SA \quad \Longrightarrow \quad \mathfrak{R}_f(T; B) \geq c_{f,\delta}\sqrt{MAT}.$$

The amplified family has $D_{\mathrm{cmp}} = \Theta(S)$. Hence the exponent $1/2$ is minimax-optimal, with a lower bound carrying the standard $\sqrt{SA}$ parallel-decision factor.

Proof. We construct the family and compute both its robust decision gap and its information divergence.

We first construct an interior Bernoulli direction with positive slope. For $q \in [0, 1]$, recall

$$y(q) := \underline{p}_f(q) = \sigma_\delta^f((q, 1-q), (1, 0)).$$

Joint convexity of the closed f-divergence shows that y is convex in q: mixing nominal rows and corresponding feasible minimizers produces a feasible mixture whose objective is the same mixture of the two objectives. Lemma 19 shows that y is continuous on $[0, 1]$. It is finite and takes values in $[0, 1]$. Moreover, $y(0) = 0$. Since $f \in \mathcal{F}_{\mathrm{ne}}$,

$$\theta_f(1) = f(0) + f'(\infty) = +\infty > \delta,$$

and therefore $y(1) > 0$. A finite convex function on $(0, 1)$ is locally Lipschitz and differentiable almost everywhere. If its derivative were nonpositive almost everywhere, convexity would make it nonincreasing on $(0, 1)$; continuity at the endpoints would then contradict $y(1) > y(0)$. Hence there is a differentiability point $q_0 \in (0, 1)$ with

$$y'(q_0) > 0.$$

The positive-slope differentiability points form a set of positive measure, so we may choose $q_0 \neq 1 - q_\star$. Avoiding this single exceptional value loses no generality. Choose constants $c_1, c_2 > 0$ and $\varepsilon_0 > 0$ so that, for every $0 < \varepsilon \leq \varepsilon_0$,

$$c_1\varepsilon \leq y(q_0 + \varepsilon) - y(q_0 - \varepsilon) \leq c_2\varepsilon. \tag{75}$$

This follows directly from the definition of $y'(q_0)$ after decreasing ε_0 if necessary. We also take $\varepsilon_0 < |1 - q_0 - q_\star|/2$. Moreover $\theta_f(q_0) > \delta$ holds automatically: $y \geq 0$, so if $y(q_0) = 0$ at an interior point of differentiability then q_0 is a local minimum and $y'(q_0) = 0$. Hence $y'(q_0) > 0$ forces $y(q_0) > 0$, that is $\theta_f(q_0) > \delta$. After another decrease of ε_0,

$$\theta_f(q_0 \pm \varepsilon) > \delta.$$

Choose $c \in (0, 1)$ with $\theta_f(c) > \delta$ and define

$$z := \max\{p : D_f((p, 1-p) \| (c, 1-c)) \leq \delta\}.$$

The nominal row is feasible, so $z \geq c > 0$.

We next define the two MDPs and compute their exact robust gains. Let the states be $\{0, 1\}$ and let the rewards be

$$r(0, a) = 0, \qquad r(1, a) = 1$$

for every action. At state 0 there are actions a, b. In model P_0, their probabilities of moving to state 1 are respectively

$$q_+ := q_0 + \varepsilon, \qquad q_- := q_0 - \varepsilon,$$

and in model P_1 these two rows are interchanged. At state 1, every action returns to state 0 with probability c in both models.

Consider a stationary policy that chooses action a at state 0 with probability $\alpha \in [0,1]$. Rectangularity lets the adversary minimize the probability of moving from 0 to the rewarding state separately in each action row. It also lets the adversary maximize the probability of leaving state 1. Thus the worst-case transition probabilities are

$$\begin{aligned} y_0(\alpha) &= \alpha y(q_+) + (1-\alpha) y(q_-), \\ y_1(\alpha) &= \alpha y(q_-) + (1-\alpha) y(q_+), \\ \widetilde{Q}(0 \mid 1) &= z. \end{aligned}$$

For a two-state chain with transition probability y from 0 to 1 and z from 1 to 0, the stationary mass on state 1 is

$$\psi(y) := \frac{y}{y+z}.$$

Consequently

$$g^{\pi_\alpha}_{\mathrm{rob},P_i} = \psi(y_i(\alpha)).$$

The function ψ is increasing and concave because

$$\psi'(y) = \frac{z}{(y+z)^2} > 0, \qquad \psi''(y) = -\frac{2z}{(y+z)^3} < 0.$$

Writing $y_+ := y(q_+)$ and $y_- := y(q_-)$, the optimal actions are a in model 0 and b in model 1, and

$$g^{\mathrm{rob}}(P_0) = g^{\mathrm{rob}}(P_1) = \psi(y_+).$$

We now prove decision separation. For every α,

$$y_0(\alpha) + y_1(\alpha) = y_+ + y_-.$$

Concavity of ψ gives

$$\psi(y_0(\alpha)) + \psi(y_1(\alpha)) \le 2\psi\left(\frac{y_+ + y_-}{2}\right).$$

Hence

$$\ell_0(\pi_\alpha) + \ell_1(\pi_\alpha) \ge 2\left[\psi(y_+) - \psi\left(\frac{y_+ + y_-}{2}\right)\right].$$

All values $y_\pm$ lie in a fixed compact subinterval of $(0,1]$ for small ε. Therefore ψ' has a strictly positive lower bound m_ψ on the interval between them. Integrating ψ' from y_- to y_+ and applying equation (75) yields

$$\begin{aligned} \Delta_{\mathrm{dec}}(P_0, P_1) &\ge m_\psi (y_+ - y_-) \\ &\ge m_\psi c_1 \varepsilon. \end{aligned}$$

Finally, we bound the information divergence and choose the separation. The two models differ only in the two Bernoulli rows at state 0. For $q_\pm$ in a fixed compact subset of $(0,1)$, the second derivatives of both displayed Bernoulli relative entropies are uniformly bounded, while their value and first derivative vanish on the diagonal. Hence

$$\max\{\mathrm{kl}(q_+\|q_-), \mathrm{kl}(q_-\|q_+)\} \le C_{\mathrm{kl}} \varepsilon^2$$

for a constant C_{kl}. The chain rule for relative entropy therefore gives, for every $\mathfrak{A} \in \mathfrak{A}_{\mathrm{pol}}$,

$$\mathrm{KL}(\mathbb{P}_0^{\mathfrak{A},T} \| \mathbb{P}_1^{\mathfrak{A},T}) \le C_{\mathrm{kl}} T \varepsilon^2.$$

Choose

$$\varepsilon := \min\left\{\frac{1}{\sqrt{2C_{\mathrm{kl}}T}}, \varepsilon_0\right\}.$$

For all sufficiently large T, the first term is selected and the KL divergence is at most $1/2$. Theorem 10 now gives

$$\max_i \mathbb{E}_i \operatorname{Reg}^{(i)}_{\mathrm{pol}}(T) \ge \frac{m_\psi c_1}{4\sqrt{2C_{\mathrm{kl}}}} \sqrt{T}.$$

Finally, the edges $0 \to 1$ and $1 \to 0$ have deletion costs strictly larger than δ. They form a communicating subgraph of the robust core, so $\delta < \delta_f^*(P_i)$, and $C(P_i) = \mathcal{S}$. All probabilities vary in fixed compact subsets of the relative interior, so $q_\pm \in [q_0 - \varepsilon_0, q_0 + \varepsilon_0] \subset (0,1)$ and $c \in (0,1)$ are bounded away from 0 and 1 uniformly in T; every positive entry of P_i is one of $q_\pm$, $1 - q_\pm$, c, $1 - c$, whence $p_{\min,C}(P_i)$ is bounded below uniformly in T. The core inequalities $\theta_f(q_\pm) > \delta$ and $\theta_f(c) > \delta$ hold with a fixed strict margin, and the choice $\varepsilon_0 < |1 - q_0 - q_\star|/2$ keeps every complementary coordinate $1 - q_\pm$ at a fixed strict distance from the deletion boundary $q_\star$, so every core-edge deletion inequality involving $1 - q_\pm$ also has a fixed strict margin. Hence $\Delta_\theta(P_i)$ is bounded below uniformly in T. The two-state bias and local row moduli are uniformly bounded. The relevant Bernoulli rows range over fixed compact interior intervals, on which the finite convex row-value functions are uniformly Lipschitz; hence the modulus definition gives a uniform bound on K_{face}. Because every positive transition on these two-state chains is uniformly bounded away from zero, $D_0(P_i)$ and $D_{\mathrm{cmp}}((P_i)_R)$ are uniformly bounded; compact Bellman stability also bounds $H_{\mathrm{loc}}((P_i)_R)$. Choosing B larger than these bounds and the reciprocal margins places both witnesses in the family $\mathfrak{P}_f(B)$. This proves the two-state assertion.

We now amplify the construction across states and actions. Let $S \geq 3$, $A \geq 2$, $M := \lfloor S/3 \rfloor$, $r := S - 3M \in \{0,1,2\}$, and $L := 2M + r$. Use decision states $x_1, \dots, x_M$, outcome states $z_1, \dots, z_M$ and $w_1, \dots, w_M$, and, when $r > 0$, connector states $c_1, \dots, c_r$. Every z_i and w_i moves deterministically to x_{i+1}, with the last pair moving through the connector chain and then to x_1. All actions at these states are identical. Rewards are one at the z_i and zero elsewhere.

Shrink ε_0 once more so that, for $q_{\mathrm{h}} := q_0 + \varepsilon$ and $q_{\mathrm{b}} := q_0$,

$$\Delta_y := y(q_{\mathrm{h}}) - y(q_{\mathrm{b}}) \geq c_0' \varepsilon, \qquad \theta_f(q_{\mathrm{h}}) \wedge \theta_f(q_{\mathrm{b}}) > \delta \tag{76}$$

for $0 < \varepsilon \leq \varepsilon_0$ and a constant $c_0' > 0$. This follows from $y'(q_0) > 0$ and the strict deletion margin already established above. Index models by $\nu = (\nu_1, \dots, \nu_M) \in [A]^M$. At x_i, action a moves to z_i with probability

$$q_{i,a}^\nu := q_{\mathrm{b}} + \varepsilon \mathbf{1}\{a = \nu_i\}$$

and to w_i otherwise. Starting at x_1, the nominal trajectory visits each decision state exactly once every L steps, independently of the chosen actions and observed outcomes. Set $s_{\mathrm{init}} := x_1$ for this lower-bound family.

For a stationary randomized policy π, the deterministic continuation from z_i and w_i makes their robust bias difference exactly one. Because $f \in \mathcal{F}_{\mathrm{ne}}$, an ambiguity row at x_i cannot leave $\{z_i, w_i\}$, and its worst-case probability of z_i is $y(q_{i,a}^\nu)$. Summing the Bellman equations around the cycle gives the exact gains

$$g_{\mathrm{rob},P_\nu}^\pi = \frac{1}{L} \sum_{i=1}^{M} \sum_{a=1}^{A} \pi(a \mid x_i) y(q_{i,a}^\nu), \tag{77}$$

$$g^{\mathrm{rob}}(P_\nu) - g_{\mathrm{rob},P_\nu}^\pi = \frac{\Delta_y}{L} \sum_{i=1}^{M} \big[1 - \pi(\nu_i \mid x_i)\big]. \tag{78}$$

In particular, the policy loss is scalar and equation (77) applies to every policy output π_t of a history-dependent policy-valued learner.

It remains to lower-bound the number of unresolved local decisions. Put a uniform prior on ν. For a fixed context i, let $\mathbb{P}_{i,0}$ be the mixture experiment in which all rows at x_i have parameter q_{b} and the remaining coordinates of ν are uniform, and let $\mathbb{P}_{i,a}$ be the analogous mixture conditional on $\nu_i = a$. Let $\eta = \nu_{-i}$, and write $\mathbb{P}_{i,b}^{t,\eta}$ for the conditional history law given $\nu_{-i} = \eta$, where $b \in \{0, a\}$. For each η, the two laws differ only in the row (x_i, a). Marginalizing η and applying the adaptive KL chain rule gives

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_{i,0}^t \| \mathbb{P}_{i,a}^t) &\leq \mathbb{E}_\eta \mathrm{KL}(\mathbb{P}_{i,0}^{t,\eta} \| \mathbb{P}_{i,a}^{t,\eta}) \\ &= \mathrm{kl}(q_{\mathrm{b}} \| q_{\mathrm{b}} + \varepsilon) \mathbb{E}_\eta \mathbb{E}_{i,0}^\eta N_{i,a}(t) \\ &\leq C_{\mathrm{kl}} \varepsilon^2 \mathbb{E}_{i,0} N_{i,a}(t), \end{aligned}$$

where $N_{i,a}(t)$ counts uses of actions a at x_i before time t.

Pinsker's inequality, followed by Cauchy–Schwarz and $\sum_a N_{i,a}(t) \le \lceil t/L \rceil$, yields

$$\frac{1}{A}\sum_{a=1}^{A} \mathbb{E}_{i,a}\pi_t(a \mid x_i) \le \frac{1}{A} + \frac{1}{A}\sum_{a=1}^{A}\sqrt{\frac{C_{\mathrm{kl}}\varepsilon^2}{2}\mathbb{E}_{i,0}N_{i,a}(t)}$$
$$\le \frac{1}{A} + \varepsilon\sqrt{\frac{C_{\mathrm{kl}}}{2A}\left\lceil \frac{t}{L} \right\rceil}. \tag{79}$$

Choose

$$\varepsilon := c_*\sqrt{\frac{LA}{T}}$$

with $c_* > 0$ small enough. For $T \ge C_{f,\delta}LA$, this lies below ε_0 and makes the right-hand side of (79) at most $3/4$ for every $t \le T$. Averaging (77) over the prior and summing the per-step regret therefore gives

$$\frac{1}{A^M}\sum_{\nu\in[A]^M} \mathbb{E}_\nu \operatorname{Reg}_{\mathrm{pol}}(T) \ge \frac{\Delta_y}{L}\,\frac{MT}{4} \ge c_{f,\delta}T\varepsilon \ge c_{f,\delta}\sqrt{SAT},$$

where M/L is bounded below by a universal positive constant and $L = \Theta(S)$. Yao's principle, or simply that a supremum dominates this Bayes average, proves the claimed minimax lower bound.

Finally, these models belong to the stated family with explicit scaling. Their nominal graphs are communicating. The edges $x_i \to z_i$ and the deterministic continuation edges form a weakly communicating subgraph of the robust core with a fixed deletion margin; the w_i are transient in that core. All positive probabilities are uniformly bounded away from zero, so $p_{\min,C}^{-1} = O_{f,\delta}(1)$. The uniform policy returns to each requested state–action pair in expected time $O_{f,\delta}(LA)$, hence $D_0 = O_{f,\delta}(SA)$. Every deterministic comparison policy cycles through the decision states, and its hitting times are $\Theta_{f,\delta}(L)$, so $D_{\mathrm{cmp}} = \Theta_{f,\delta}(S)$. The same cycle equations, on a fixed compact neighborhood of the displayed probabilities, give $H_{\mathrm{loc}} = O_{f,\delta}(S)$. Thus all models lie in $\mathfrak{P}_f(B)$ whenever $B \ge C_{f,\delta}SA$, completing the amplified claim. □

Proposition 10. *Let $f \in \mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$ and $0 < \delta < \theta_f(1)$. There are constants $B, c_{f,\delta}, T_0 > 0$ such that, for every $T \ge T_0$,*

$$\mathfrak{R}_f^{\mathrm{all}}(T;B) \ge c_{f,\delta}\sqrt{T}, \qquad \mathfrak{R}_f^{\mathrm{bell}}(T;B) \ge c_{f,\delta}\sqrt{T}.$$

The witnesses have three states and three actions and lie strictly in the two respective annuli.

Proof. Write $q_\star := q_\star(f,\delta)$. The assumptions $f \notin \mathcal{F}_{\mathrm{nc}}$ and $\delta < \theta_f(1)$ give

$$0 < q_\star < 1.$$

Use the differentiability point q_0, constants c_1, c_2, ε_0, return probability c, robust minimum $y(q) = \underline{p}_f(q)$, robust maximum return probability z, and function $\psi(u) = u/(u+z)$ constructed in the proof of Proposition 9. In particular, after decreasing ε_0 once more,

$$q_\pm := q_0 \pm \varepsilon > q_\star, \qquad c_1\varepsilon \le y(q_+) - y(q_-) \le c_2\varepsilon, \tag{80}$$

for $0 < \varepsilon \le \varepsilon_0$, and $c > q_\star$. Choose fixed numbers

$$0 < \kappa_{\mathrm{n}} < \tfrac{1}{4}\min\{q_\star, 1-q_\star\}, \tag{81}$$

$$\frac{q_\star}{2} < \kappa_{\mathrm{m}} < \min\{q_\star, \tfrac{1}{2}\}, \qquad p_{\mathrm{e}} := \frac{1+q_\star}{2} > q_\star. \tag{82}$$

The strict interval for κ_{m} is nonempty for every $q_\star \in (0,1)$.

Both pairs use states $\{0,1,w\}$, actions $\{a,b,n\}$, rewards

$$r(1,\cdot) = 1, \qquad r(0,\cdot) = r(w,\cdot) = 0,$$

and initial state 0. At state 0, models P_0, P_1 interchange the two hard rows

$$P_0(1 \mid 0,a) = q_+, \qquad P_0(1 \mid 0,b) = q_-,$$
$$P_1(1 \mid 0,a) = q_-, \qquad P_1(1 \mid 0,b) = q_+,$$

with the complementary mass on 0. The navigation row is common:

$$P_i(w \mid 0, n) = \kappa_{\mathrm{n}}, \qquad P_i(0 \mid 0, n) = 1 - \kappa_{\mathrm{n}}.$$

At state 1, every action returns to 0 with probability c and stays at 1 otherwise. These rows are also common to the two models.

We now give two choices for the rows at w.

Memberwise-communicating scaffold. For action a, consider

$$P_i(0 \mid w, a) = P_i(1 \mid w, a) = \kappa_{\mathrm{m}}, \qquad P_i(w \mid w, a) = 1 - 2\kappa_{\mathrm{m}}. \tag{83}$$

For actions b, n,

$$P_i(0 \mid w, b) = P_i(0 \mid w, n) = p_{\mathrm{e}}, \qquad P_i(w \mid w, b) = P_i(w \mid w, n) = 1 - p_{\mathrm{e}}. \tag{84}$$

Each of the two individual exit edges in (83) is deletable, but their total mass is not:

$$\kappa_{\mathrm{m}} \le q_\star < 2\kappa_{\mathrm{m}}.$$

Thus the robust core row for (w, a) has no exit from w, or is empty, whereas every feasible realization of that row retains some exit to $\{0, 1\}$. The core contains the undeletable edges $0 \to 1$, $1 \to 0$, and the escape edge $w \to 0$, but is not weakly communicating because the fragile (w, a) row either supplies a separate self-loop recurrent class or is empty. Hence $\delta_f^*(P_i) \le \delta$.

On the other hand, every feasible support graph is weakly communicating. The states $0, 1$ communicate through the undeletable hard edges. Every action at w retains an exit to this class. If the navigation edge $0 \to w$ is deleted, w is transient; if it is retained, it joins the same communicating class. Theorem 2 therefore gives

$$\delta < \delta_f^\dagger(P_i).$$

Multichain-admitting scaffold. Replace (83) by

$$P_i(0 \mid w, a) = \kappa_{\mathrm{n}}, \qquad P_i(w \mid w, a) = 1 - \kappa_{\mathrm{n}}, \tag{85}$$

and keep the escape rows (84). The set $I = \{0, 1\}$ is δ-closed: its only nominal exit is the navigation mass $\kappa_{\mathrm{n}} \le q_\star$. The disjoint singleton $J = \{w\}$ is δ-closable through action a. Consequently

$$\delta_f^*(P_i) \le \delta_f^\dagger(P_i) \le \delta.$$

Both scaffolds satisfy $\delta < \delta_f^{\mathrm{bell}}(P_i)$. This can be checked without continuity. In either scaffold the only proper nominally closable sets are $\{0, 1\}$ and $\{0, w\}$. Their respective disjoint complements $\{w\}$ and $\{1\}$ are not δ-closed: in the memberwise scaffold the former has aggregate exit mass $2\kappa_{\mathrm{m}} > q_\star$, in the multichain scaffold it has escape mass $p_{\mathrm{e}} > q_\star$, and in both scaffolds the latter has return mass $c > q_\star$. The remaining proper subsets are not nominally closable. Equation (4) now yields

$$\delta < \delta_f^{\mathrm{bell}}(P_i)$$

with a fixed positive deletion-mass margin in both constructions.

It remains to prove the learning lower bound. Fix an arbitrary stationary policy π and write

$$\alpha_a := \pi(a \mid 0), \qquad \alpha_b := \pi(b \mid 0), \qquad \alpha_n := 1 - \alpha_a - \alpha_b.$$

Set

$$u_0 := \alpha_a y(q_+) + \alpha_b y(q_-),$$
$$u_1 := \alpha_a y(q_-) + \alpha_b y(q_+).$$

For the hard rows, the adversary uses the minimizing Bernoulli rows from Proposition 9; at state 1 it uses the row attaining the maximum return probability z. On the navigation row it deletes the mass κ_{n} on w and stays at 0, which is feasible because $\kappa_{\mathrm{n}} \le q_\star$. This supplies, from states 0 and 1, a feasible adversarial kernel with average gain $\psi(u_i)$ in model i. Hence

$$g^\pi_{\mathrm{rob},P_i}(x) \le \psi(u_i), \qquad x \in \{0, 1\}. \tag{86}$$

Starting from w, the same upper bound holds whenever the announced policy assigns positive probability to an escape that the adversary cannot delete: the transient prefix at w does not affect average gain. In the multichain scaffold, if the policy instead uses only action a at w, the adversary closes $\{w\}$ and the robust gain from w is zero. In the memberwise scaffold every action has a nondeletable aggregate exit, so (86) holds at w as well. Thus, in all cases, either (86) holds or the policy has zero gain from w in both models.

The hard policy that chooses the better of a, b at 0 and an escape at w has scalar robust gain $\psi(y(q_+))$. The preceding adversarial upper bound shows that no policy can do better. Therefore

$$g^{\mathrm{rob}}(P_0) = g^{\mathrm{rob}}(P_1) = \psi(y(q_+)).$$

Moreover,

$$u_0 + u_1 = (\alpha_a + \alpha_b)[y(q_+) + y(q_-)] \le y(q_+) + y(q_-).$$

Since ψ is increasing and concave, equation (86) gives, uniformly in the current state,

$$\begin{aligned}\ell_0(\pi, x) + \ell_1(\pi, x) &\ge 2\psi(y(q_+)) - \psi(u_0) - \psi(u_1) \\ &\ge 2\left[\psi(y(q_+)) - \psi\left(\frac{y(q_+) + y(q_-)}{2}\right)\right] \\ &\ge c_{\mathrm{dec}}\varepsilon.\end{aligned}$$

When the policy has zero gain from w, the same inequality is immediate after decreasing c_{dec} if necessary. Hence

$$\Delta^{\mathrm{st}}_{\mathrm{dec}}(P_0, P_1) \ge c_{\mathrm{dec}}\varepsilon. \tag{87}$$

The models differ only in the two Bernoulli hard rows at state 0. Exactly as in Proposition 9, the adaptive chain rule gives

$$\mathrm{KL}(\mathbb{P}_0^{\mathfrak{A},T} \| \mathbb{P}_1^{\mathfrak{A},T}) \le C_{\mathrm{kl}} T \varepsilon^2.$$

Choosing $\varepsilon = (2C_{\mathrm{kl}}T)^{-1/2} \wedge \varepsilon_0$ and applying Theorem 11 to (87) proves an $\Omega_{f,\delta}(\sqrt{T})$ lower bound in each annulus.

Finally, all nonzero probabilities other than $q_\pm$ are fixed, and $q_\pm$ range in a compact interior interval with a strict deletion margin. Equations (81) give fixed strict margins for the two structural classifications and for Δ_{bell}. The nominal control graphs are communicating and have $C(P_i) = \mathcal{S}$: the hard actions make $0, 1$ recurrent, while the navigation and escape actions make $0, w$ recurrent under another deterministic policy. Thus $p^{-1}_{\min,C}$ and D_0 are uniformly bounded. A comparison policy uses a hard action at 0 and an escape at w; the finite comparison-policy class then gives a uniform D_{cmp} bound. Fixed-face Lipschitzness and compact Bellman stability give uniform bounds on K_{face} and H_{loc}. Taking B larger than these finitely many constants puts the first pair in $\mathfrak{P}_f^{\mathrm{all}}(B)$ and the second in $\mathfrak{P}_f^{\mathrm{bell}}(B)$ for every sufficiently large T. □

Theorem 14. *Let $f \in \mathcal{F} \setminus \mathcal{F}_{\mathrm{ne}}$ and $0 < \delta < \theta_f(1)$. There are $B, c_{f,\delta}, T_0 > 0$ such that*

$$c_{f,\delta} T^{2/3} \;\le\; \mathfrak{R}_f(T; B), \qquad \mathfrak{R}_f(T; B) = \widetilde{O}_{B,S,A,f,\delta}(T^{2/3}), \qquad T \ge T_0.$$

The lower bound is witnessed for each T by a pair $P_{0,T}, P_{1,T} \in \mathfrak{P}_f(B)$ on $S = 2$ differing only in two interior rows; the upper bound is Theorem 9, whose $T^{2/3}$ leading instance factors involve only $H_{\mathrm{loc}}(P_R)$, $K_{\mathrm{face}}(P_R)$, $D_0(P)$, d_R and A. On $\mathfrak{P}_f(B)$ the first and third are at most B, $K_{\mathrm{face}} \le 2B$, and d_R, A are fixed. The exponent over the whole family is therefore $2/3$ up to logarithmic factors, and no uniform square-root theorem is possible for either creation class, even under $\delta < \delta^*_f(P)$.

Proof. We again give the construction, its exact robust gains, the information bound, and the upper bound up to logarithmic factors. Fix an arbitrary $\mathfrak{A} \in \mathfrak{A}_{\mathrm{pol}}$ throughout the lower-bound argument.

We first construct two fixed robust transition levels. Use the same convex function

$$y(q) := \underline{p}_f(q).$$

We have $y(0) = 0$. The strict inequality $\delta < \theta_f(1)$ implies $y(1) > 0$. Applying the convexity/differentiability argument in the proof of Proposition 9, there are $q_0 \in (0, 1)$ with $q_0 \ne 1 - q_\star$ and $c_1, c_2, \varepsilon_0 > 0$ with $\varepsilon_0 < |1 - q_0 - q_\star|/2$, such that, for $q_\pm = q_0 \pm \varepsilon$ and $0 < \varepsilon \le \varepsilon_0$,

$$c_1\varepsilon \le y(q_+) - y(q_-) \le c_2\varepsilon, \tag{88}$$

and $\theta_f(q_\pm) > \delta$ with a strict margin. The choice of q_0 keeps every complementary coordinate $1 - q_\pm$ at a fixed strict distance from the deletion boundary $q_\star$.

Because $f \notin \mathcal{F}_{\mathrm{ne}}$, $f'(\infty) < +\infty$, so a row centered at a point mass can create a previously absent edge at finite divergence cost. Define

$$\bar{x} := \max\{x \in [0,1] : D_f((1-x,x)\|(1,0)) \le \delta\}.$$

Continuity of the divergence at $x = 0$ and $\delta > 0$ give $\bar{x} > 0$. Moving all mass to the second coordinate has cost $\theta_f(1)$, so $\delta < \theta_f(1)$ gives $\bar{x} < 1$. Thus

$$0 < \bar{x} < 1.$$

We next define the revealing MDPs. Let the states be $\{s, z\}$. At s there are a default action d and a revealing action e, with

$$\begin{aligned} r(s,d) &= 1, & P(s \mid s,d) &= 1,\\ r(s,e) &= 0, & P(z \mid s,e) &= 1. \end{aligned}$$

At z there are two zero-reward actions a, b. In model P_0, their probabilities of returning to s are q_+, q_-, respectively, and in model P_1 the two rows are interchanged. All unspecified rewards are zero. The global action set is a fixed two-element set, and d, e and a, b are state-dependent descriptive names for its elements, with d and a naming the same action and e and b naming the same action. Set $s_{\mathrm{init}} := s$.

Consider a stationary randomized policy that chooses e at s with probability α and action a at z with probability β. Under rectangular ambiguity, the worst-case probability of moving from s to z is

$$x(\alpha) = (1-\alpha)\bar{x} + \alpha,$$

because the default row can create mass $\bar{x}$ on z, whereas the nominal revealing row itself attains probability one. The worst-case return probabilities from z are

$$\begin{aligned} y_0(\beta) &= \beta y(q_+) + (1-\beta) y(q_-),\\ y_1(\beta) &= \beta y(q_-) + (1-\beta) y(q_+). \end{aligned}$$

The expected one-step reward at s is $1 - \alpha$. Hence the exact robust gain is

$$G_i(\alpha,\beta) := g^{\pi_{\alpha,\beta}}_{\mathrm{rob},P_i} = \frac{(1-\alpha) y_i(\beta)}{x(\alpha) + y_i(\beta)}. \tag{89}$$

For fixed $y > 0$, the right side decreases with α, and for fixed α it increases with y. Therefore the optimal policy has $\alpha = 0$ and chooses the action with robust return probability $y_+ := y(q_+)$. Writing $y_- := y(q_-)$,

$$g^{\mathrm{rob}}(P_0) = g^{\mathrm{rob}}(P_1) = \psi(y_+), \qquad \psi(y) := \frac{y}{\bar{x} + y}.$$

We now establish decision separation and the price of revealing. For fixed α, define

$$\psi_\alpha(y) := \frac{(1-\alpha) y}{x(\alpha) + y}.$$

This function is increasing and concave in y, and $\psi_\alpha(y) \le \psi_0(y) = \psi(y)$. Since $y_0(\beta) + y_1(\beta) = y_+ + y_-$, concavity gives

$$\begin{aligned} G_0(\alpha,\beta) + G_1(\alpha,\beta) &\le 2\psi_\alpha\left(\frac{y_+ + y_-}{2}\right)\\ &\le 2\psi\left(\frac{y_+ + y_-}{2}\right). \end{aligned}$$

Consequently, for every stationary policy,

$$\begin{aligned} \ell_0(\pi_{\alpha,\beta}) + \ell_1(\pi_{\alpha,\beta}) &\ge 2\left[\psi(y_+) - \psi\left(\frac{y_+ + y_-}{2}\right)\right]\\ &\ge c_{\mathrm{dec}}\varepsilon. \end{aligned} \tag{90}$$

Inequality (90) follows by integrating the positive minimum of ψ' on a fixed compact interval and applying equation (88). Thus

$$\Delta_{\mathrm{dec}}(P_0, P_1) \geq c_{\mathrm{dec}}\varepsilon.$$

Revealing itself has a fixed first-order cost. Even if the policy always uses the better action at z, its gain is at most

$$\frac{(1-\alpha)y_+}{\bar{x} + \alpha(1-\bar{x}) + y_+}.$$

Direct subtraction gives

$$\begin{aligned}
&\frac{y_+}{\bar{x}+y_+} - \frac{(1-\alpha)y_+}{\bar{x}+\alpha(1-\bar{x})+y_+} \\
&\quad = \frac{\alpha y_+(1+y_+)}{(\bar{x}+y_+)[\bar{x}+\alpha(1-\bar{x})+y_+]} \\
&\quad \geq \alpha \frac{y_+}{\bar{x}+y_+} \geq c_{\mathrm{rev}}\alpha
\end{aligned}$$

for a constant $c_{\mathrm{rev}} > 0$ uniform in small ε.

We next prove that the information divergence is proportional to the number of reveals. Let $N_e(T)$ be the number of times the learner actually selects e at s, and let $N_z(T)$ be the number of transitions observed from z. The nominal default action never enters z, while every revealing action enters z once. Let

$$q_{\min} := q_0 - \varepsilon_0 > 0.$$

At every visit to z, under either action and either model, the conditional probability of returning to s is at least $q_{\min}$. Therefore each visit initiated by a reveal has expected length at most $1/q_{\min}$. Conditioning successively on the reveal times gives

$$\mathbb{E}_0[N_z(T)] \leq \frac{1}{q_{\min}} \mathbb{E}_0[N_e(T)]. \tag{91}$$

For a general policy-valued learner, let $\alpha_t := \pi_t(e \mid s)$. The preceding calculation applies to the stationary policy output at time t and gives $\ell_0(\pi_t) \geq c_{\mathrm{rev}}\alpha_t$. Conditional on the history, an actual reveal at time t has probability $\mathbf{1}\{S_t = s\}\alpha_t \leq \alpha_t$. Hence

$$\mathbb{E}_0[N_e(T)] \leq \mathbb{E}_0\left[\sum_{t<T} \alpha_t\right].$$

Consequently

$$\mathbb{E}_0 \operatorname{Reg}^{(0)}_{\mathrm{pol}}(T) \geq c_{\mathrm{rev}} \mathbb{E}_0[N_e(T)]. \tag{92}$$

The models differ only in the transition rows observed at z. The bounded-second-derivative calculation in the proof of Proposition 9, together with compactness of $q_\pm$, gives

$$\max\{\mathrm{kl}(q_+ \| q_-), \mathrm{kl}(q_- \| q_+)\} \leq C_{\mathrm{kl}}\varepsilon^2.$$

The adaptive chain rule for relative entropy and equations (91) now yield

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0^{\mathfrak{A},T} \| \mathbb{P}_1^{\mathfrak{A},T}) &\leq C_{\mathrm{kl}}\varepsilon^2 \mathbb{E}_0[N_z(T)] && \text{by the adaptive KL chain rule} \\
&\leq \frac{C_{\mathrm{kl}}}{q_{\min}} \varepsilon^2 \mathbb{E}_0[N_e(T)] && \text{by (91).}
\end{aligned} \tag{93}$$

We now derive the $T^{2/3}$ lower bound. Choose

$$\varepsilon = c_0 T^{-1/3},$$

where $c_0 > 0$ is small enough that $\varepsilon \leq \varepsilon_0$ and the constant below is at most $1/2$. There are two cases.

If

$$\mathbb{E}_0[N_e(T)] \geq b_0 T^{2/3}$$

for a fixed $b_0 > 0$, then inequality (92) gives

$$\mathbb{E}_0 \operatorname{Reg}_{\mathrm{pol}}^{(0)}(T) \geq c_{\mathrm{rev}} b_0 T^{2/3}.$$

Otherwise inequality (93) gives

$$\mathrm{KL}(\mathbb{P}_0^{\mathfrak{A},T} \| \mathbb{P}_1^{\mathfrak{A},T}) \leq \frac{C_{\mathrm{kl}} b_0 c_0^2}{q_{\min}} \leq \frac{1}{2}.$$

Theorem 10 and the decision separation then give

$$\max_i \mathbb{E}_i \operatorname{Reg}_{\mathrm{pol}}^{(i)}(T) \geq \frac{c_{\mathrm{dec}} c_0}{4} T^{2/3}.$$

Taking the smaller of the two positive constants proves the lower bound.

Finally, we bound the structural constants and prove the upper bound up to logarithmic factors. The nominal graph is communicating: action e reaches z from s and both diagnostic actions return to s with positive probability. The robust core contains the default self-loop, the revealing edge $s \to z$, and both edges $z \to s$. Their deletion costs exceed δ by fixed strict margins. Hence

$$\delta < \delta_f^*(P_i) \leq \delta_f^\dagger(P_i), \qquad C(P_i) = \mathcal{S}.$$

All nonvarying transition levels are fixed, $q_\pm$ remain in a compact interior interval, and $\bar{x}, y_\pm$ stay bounded away from zero and the degenerate endpoints. Every positive entry of $P_{i,T}$ is one of 1, $q_\pm$ or $1 - q_\pm$, so $p_{\min,C}(P_{i,T})$ is bounded below uniformly in T, and every core-edge deletion inequality, including those involving the complementary coordinates $1 - q_\pm$, holds with a fixed strict margin by the choice $\varepsilon_0 < |1 - q_0 - q_\star|/2$. So $\Delta_\theta(P_{i,T})$ is bounded below uniformly in T. The uniformly positive two-state transitions bound $D_0(P_{i,T})$, while compact Bellman stability bounds $H_{\mathrm{loc}}(P_{i,T})$, uniformly in T. For the fixed-face constant, each deterministic row lies on a singleton face and therefore contributes zero, while the rows containing $q_\pm$ remain on a fixed compact subset of the Bernoulli interior. The convex interior estimate in Lemma 15 therefore gives $K_{\mathrm{face}}(P_R) < \infty$ uniformly in T. Taking B larger than all of these finitely many constants gives $P_{0,T}, P_{1,T} \in \mathfrak{P}_f(B)$ for every T.

For the upper bound, here $f \notin \mathcal{F}_{\mathrm{ne}}$, so every $P \in \mathfrak{P}_f(B)$ satisfies $C(P) = \mathcal{S}$ and $\delta < \delta_f^*(P) \leq \delta_f^\dagger(P) \leq \delta_f^{\mathrm{bell}}(P)$, so the structural and Bellman hypotheses of Theorem 9 hold, and the third is vacuous because $f \notin \mathcal{F}_{\mathrm{ne}}$. The remaining neighborhood and support-recovery hypotheses hold for all $T \geq T_0$ with T_0 depending only on B, S, A, f, δ and α: by Lemma 9(ii)–(iii),

$$\bar{\epsilon}_{\mathrm{face}}(P_R) \geq \tfrac{1}{4} \min\Big\{ \tfrac{1}{2}, \eta_f, \tfrac{1}{L_f B}, \tfrac{1}{B} \Big\} > 0$$

uniformly on the family, whereas Lemma 17 and $D_0(P) \leq B$, $d_R = S$ give $\eta_T = \widetilde{O}(\sqrt{BSA}\, T^{-1/3}) \to 0$ and $n_T(P, \alpha) = \widetilde{\Omega}(T^{2/3}/(BSA))$, whereas $n_{\mathrm{supp}}(P, \alpha) = \widetilde{O}(B)$. These bounds are uniform on the family. On the complementary range $T < T_0$ the trivial bound $\operatorname{Reg}_{\mathrm{pol}}(T) \leq T \leq T_0$ applies. Hence the $T^{2/3}$ display of Theorem 9 holds uniformly on $\mathfrak{P}_f(B)$ with $K_{\mathrm{face}}(P_R) \leq 2B$, $H_{\mathrm{loc}}(P_R) \leq B$ and $D_0(P) \leq B$, and $\mathfrak{R}_f(T; B) = \widetilde{O}_{B,S,A,f,\delta}(T^{2/3})$ for $T \geq T_0$. With the lower bound this gives the stated $\widetilde{\Theta}(T^{2/3})$ rate.

The two-model subproblem attains the same exponent directly, without the parameter bounds. Suppose the learner is told that the instance is one of these two models. It can reveal until it has collected $n = \lceil T^{2/3} \rceil$ diagnostic transitions, splitting them evenly between a and b, and then choose the model with larger likelihood. The expected reveal duration is at most $n/q_{\min}$, so its exploration regret is $O(n)$. A standard likelihood-ratio test has error probability at most a constant because $n\varepsilon^2 = \Theta(1)$. On an incorrect decision, the committed policy loses $O(\varepsilon)$ per step by equation (88) and the Lipschitz continuity of ψ. Thus its expected commitment regret is

$$O(T\varepsilon) = O(T^{2/3}).$$

The exploration and commitment bounds give the matching $O(T^{2/3})$ upper bound on this subproblem. □

Proposition 11. *Let $f \in \mathcal{F} \setminus (\mathcal{F}_{\mathrm{ne}} \cup \mathcal{F}_{\mathrm{nc}})$ and $0 < \delta < \theta_f(1)$. There are $B, c_{f,\delta}, T_0 > 0$ such that*

$$c_{f,\delta} T^{2/3} \le \mathfrak{R}_f^{\mathrm{bell}}(T; B) \le \widetilde{O}_{B,f,\delta,S,A}(T^{2/3})$$

for every $T \ge T_0$.

Proof. We modify the revealing pair of Theorem 14 by one fixed, model-independent row that places the pair above $\delta_f^\dagger$. Write $q_\star := q_\star(f, \delta)$. The assumptions $f \notin \mathcal{F}_{\mathrm{nc}}$ and $\delta < \theta_f(1)$ imply

$$0 < q_\star < 1.$$

Use the differentiability point q_0, the robust Bernoulli minimum $y(q) = \underline{p}_f(q)$, and $q_\pm = q_0 \pm \varepsilon$ from the proof of Theorem 14. We choose that point away from the single value $1 - q_\star$ and decrease the fixed radius ε_0 so that, for $0 < \varepsilon \le \varepsilon_0$,

$$q_\pm > q_\star, \qquad c_1 \varepsilon \le y(q_+) - y(q_-) \le c_2 \varepsilon, \tag{94}$$

and every core-edge deletion inequality involving $q_\pm$ or $1 - q_\pm$ has a fixed strict margin whenever that edge is in the core. Positive-slope differentiability points occupy a set containing more than one point, so avoiding the one exceptional value loses no generality. Choose fixed

$$0 < \kappa < \tfrac{1}{4} \min\{q_\star, 1 - q_\star\}$$

and a fixed $q_{\mathrm{l}} \in (q_\star, q_-)$ such that $y(q_{\mathrm{l}}) < y(q_-)$ and neither q_{l} nor $1 - q_{\mathrm{l}}$ lies on a deletion boundary. Such a choice follows by taking q_{l} sufficiently close to, but strictly to the left of, the positive-slope point q_0.

Let the states be $\{s, z\}$ and the common action set be $\{d, e, a, b, c\}$. The only positive reward is

$$r(s, d) = 1.$$

At s, action d is the point mass at s, while every action other than d has the common diagnostic row

$$P_i(z \mid s, u) = \kappa, \qquad P_i(s \mid s, u) = 1 - \kappa, \qquad u \ne d.$$

At z, actions a, b have return probabilities to s equal to q_+, q_- in model 0 and in the opposite order in model 1. Actions d, e return with the common probability q_{l}. The remaining action c has

$$P_i(s \mid z, c) = \kappa, \qquad P_i(z \mid z, c) = 1 - \kappa.$$

All unspecified probability remains at the current state.

We first locate the pair structurally. Both nominal graphs are communicating. The singleton $I = \{s\}$ is δ-closed: every nominal exit from it has mass $\kappa \le q_\star$ and can be deleted. The disjoint singleton $J = \{z\}$ is δ-closable through action c, whose return mass is also κ. The exact formula for $\delta_f^\dagger$ therefore gives

$$\delta_f^*(P_i) \le \delta_f^\dagger(P_i) \le \delta.$$

Since $f \notin \mathcal{F}_{\mathrm{ne}}$, Theorem 3 gives $\delta_f^{\mathrm{bell}}(P_i) = +\infty$.

We next compute robust gains. Let

$$\bar{x} := \max\{x : D_f((1 - x, x) \| (1, 0)) \le \delta\}$$

as in Theorem 14, and let

$$x_\kappa := \max\{x : D_f((1 - x, x) \| (1 - \kappa, \kappa)) \le \delta\}.$$

Here $\bar{x}, x_\kappa > 0$. For a stationary policy write $\lambda := \pi(d \mid s)$ and let $\beta_u := \pi(u \mid z)$. Rectangularity and monotonicity of the two-state average reward give the exact robust return probability

$$\begin{aligned} y_0(\beta) &:= \beta_a y(q_+) + \beta_b y(q_-) + (\beta_d + \beta_e) y(q_{\mathrm{l}}), \\ y_1(\beta) &:= \beta_a y(q_-) + \beta_b y(q_+) + (\beta_d + \beta_e) y(q_{\mathrm{l}}), \end{aligned}$$

where action c contributes zero because its entire return mass is deletable. Thus, with the continuous convention that the expression is zero when $y_i(\beta) = 0$,

$$G_i(\lambda, \beta) := g^\pi_{\mathrm{rob}, P_i} = \frac{\lambda y_i(\beta)}{\lambda \bar{x} + (1 - \lambda) x_\kappa + y_i(\beta)}. \tag{95}$$

The gain is scalar also in the zero-return cases: the adversary then makes z absorbing and reaches it from s with positive probability.

For fixed $y > 0$, the right side of (95) is increasing in λ; for fixed λ it is increasing in y. Hence the optimal policy chooses d at s and the better hard action at z, and

$$g^{\mathrm{rob}}(P_i) = \psi(y_+), \qquad \psi(y) := \frac{y}{\bar{x} + y}, \qquad y_+ := y(q_+).$$

Moreover

$$y_0(\beta) + y_1(\beta) \le y(q_+) + y(q_-),$$

because $2y(q_l) < 2y(q_-) \le y(q_+) + y(q_-)$. Since $G_i(\lambda, \beta) \le \psi(y_i(\beta))$ and ψ is increasing and concave, the calculation in (90) gives

$$\ell_0(\pi) + \ell_1(\pi) \ge c_{\mathrm{dec}}\varepsilon. \tag{96}$$

Using $y_i(\beta) \le y_+$ in (95), direct subtraction also gives the uniform diagnostic cost

$$\ell_i(\pi) \ge \psi(y_+) - \frac{\lambda y_+}{\lambda\bar{x} + (1-\lambda)x_\kappa + y_+} \tag{97}$$

$$= \frac{(1-\lambda)y_+(x_\kappa + y_+)}{(\bar{x} + y_+)[\lambda\bar{x} + (1-\lambda)x_\kappa + y_+]} \ge c_{\mathrm{rev}}(1-\lambda), \tag{98}$$

where $c_{\mathrm{rev}} > 0$ is uniform for small ε.

Let $N_D(T)$ count nondefault actions actually selected at s, and let $N_z^{a,b}(T)$ count informative a, b transitions from z. Conditional on the history, the former occurs with probability $\mathbf{1}\{S_t = s\}[1-\lambda_t] \le 1 - \lambda_t$. Hence (97) implies

$$\mathbb{E}_i \operatorname{Reg}_{\mathrm{pol}}^{(i)}(T) \ge c_{\mathrm{rev}}\mathbb{E}_i N_D(T).$$

Under the nominal dynamics, every action at z returns to s with probability at least

$$r_{\min} := \min\{\kappa, q_l, q_0 - \varepsilon_0\} > 0.$$

Each visit to z is initiated by a nondefault action at s, so the same geometric-excursion argument as in (91) yields

$$\mathbb{E}_0 N_z^{a,b}(T) \le r_{\min}^{-1}\mathbb{E}_0 N_D(T).$$

The models differ only in the two informative rows. Therefore

$$\mathrm{KL}(\mathbb{P}_0^{\mathfrak{A},T} \| \mathbb{P}_1^{\mathfrak{A},T}) \le C_{\mathrm{kl}} r_{\min}^{-1}\varepsilon^2 \mathbb{E}_0 N_D(T). \tag{99}$$

The two-case argument following (93), with $\varepsilon = c_0 T^{-1/3}$, now combines (96)– (99) to give

$$\max_i \mathbb{E}_i \operatorname{Reg}_{\mathrm{pol}}^{(i)}(T) \ge cT^{2/3}.$$

Finally, all nonvarying positive probabilities are fixed and $q_\pm$ stay in a compact interior interval. The choices above keep every surviving core edge a fixed distance from the deletion boundary, so Δ_θ and $p_{\min,C}$ are uniformly positive. Uniform exploration has bounded hitting times because κ is fixed, and compact Bellman stability gives bounded H_{loc}. In the creation class $m_{\mathrm{bell}} = +\infty$, so Δ_{bell} is bounded below by the same positive probability margin. Taking B above these constants places both models in $\mathfrak{P}_f^{\mathrm{bell}}(B)$. The upper bound is Theorem 9, completing the proof. □

Proposition 12. *Let $f \in \mathcal{F} \setminus (\mathcal{F}_{\mathrm{ne}} \cup \mathcal{F}_{\mathrm{nc}})$ and $0 < \delta < \theta_f(1)$. There are $B, c_{f,\delta}, T_0 > 0$ such that*

$$c_{f,\delta}T^{2/3} \le \mathfrak{R}_f^{\mathrm{all}}(T; B) \le \widetilde{O}_{B,f,\delta,S,A}(T^{2/3})$$

for every $T \ge T_0$.

Proof. We place the revealing pair of Theorem 14 above the common-core radius without creating a multichain ambiguity member. Write $q_\star := q_\star(f, \delta)$. Because deletion and creation are both possible and $\delta < \theta_f(1)$,

$$0 < q_\star < 1.$$

Let $y(q) := \underline{p}_f(q)$. As in the proof of Proposition 11, choose a positive-slope point $q_0 > q_\star$, a fixed $q_1 \in (q_\star, q_0)$, and $\varepsilon_0, c_1, c_2 > 0$ so that, for $q_\pm = q_0 \pm \varepsilon$ and $0 < \varepsilon \le \varepsilon_0$,

$$q_1 < q_- < q_+, \qquad c_1\varepsilon \le y(q_+) - y(q_-) \le c_2\varepsilon, \qquad y(q_1) < y(q_-), \tag{100}$$

and all these probabilities and their complements stay a fixed distance from every deletion boundary that they do not cross. Choose fixed numbers

$$0 < \kappa_{\mathrm{n}} < \tfrac{1}{4}\min\{q_\star, 1 - q_\star\}, \qquad \tfrac{1}{2}q_\star < \kappa_{\mathrm{m}} < \min\{q_\star, \tfrac{1}{2}\}. \tag{101}$$

If necessary perturb κ_{m} within this open interval so that $1 - 2\kappa_{\mathrm{m}}$ is not on a deletion boundary.

Use states $\{s, z, w\}$ and the common action set $\{d, e, a, b, c, n\}$. The rewards are

$$r(s, d) = 1, \qquad r(w, u) = 1 \quad (u \in \mathcal{A}), \qquad r(s, u) = 0 \quad (u \ne d), \qquad r(z, u) = 0 \quad (u \in \mathcal{A}).$$

At s, action d is the point mass at s, action e and every action in $\{a, b, c\}$ are point masses at z, and

$$P_i(w \mid s, n) = \kappa_{\mathrm{n}}, \qquad P_i(s \mid s, n) = 1 - \kappa_{\mathrm{n}}.$$

At z, actions a, b have return probabilities to s equal to q_+, q_- in model 0 and in the opposite order in model 1. Every other action at z returns to s with probability q_1; the remaining mass stays at z. At w, consider

$$P_i(s \mid w, c) = P_i(z \mid w, c) = \kappa_{\mathrm{m}}, \qquad P_i(w \mid w, c) = 1 - 2\kappa_{\mathrm{m}}, \tag{102}$$

while every action $u \ne c$ is the point mass at s. The two models differ only in the rows (z, a) and (z, b).

We first verify the strict structural location. The nominal graphs are communicating. In the robust core, the edges $s \to z$ and $z \to s$ survive with fixed margins. In the row (w, c), however, the two exits to s and z are individually deletable, whereas their aggregate is not:

$$\kappa_{\mathrm{m}} \le q_\star < 2\kappa_{\mathrm{m}}.$$

Thus the intersection core either retains a separate self-loop at w or has an empty (w, c) row, and in either case it is not weakly communicating. Theorem 1 gives $\delta_f^*(P_i) \le \delta$.

On the other hand every feasible ambiguity support is weakly communicating. The states s, z communicate through undeletable edges. Every feasible realization of every action at w retains an exit to $\{s, z\}$: this is immediate for the point-mass rows, and for action c it follows from the undeletable aggregate mass $2\kappa_{\mathrm{m}}$. If the fragile navigation edge $s \to w$ is deleted, w is transient under every deterministic policy; if it remains, w joins the same communicating class under a suitable policy. Hence Theorem 2 gives

$$\delta < \delta_f^\dagger(P_i).$$

Creation makes $\delta_f^{\mathrm{bell}}(P_i) = +\infty$, so the pair lies strictly in the claimed annulus.

We next prove that the scaffold preserves the revealing control problem. Let

$$\bar{x} := \max\{x : D_f((1 - x, x) \| (1, 0)) \le \delta\}, \qquad \psi(y) := \frac{y}{\bar{x} + y}.$$

As in Theorem 14, $0 < \bar{x} < 1$. Consider the policy that uses d at s and at w, and the better hard action at z. Aggregate $H := \{s, w\}$ into one reward-one state and let $L := \{z\}$. Data processing for f-divergence shows that every ambiguity row from a point mass in H has probability at most $\bar{x}$ of entering L, and every ambiguity row centered at the hard z-row has probability at least $y(q_+)$ of returning to H. Stationary flow between H and L in each recurrent class therefore gives reward at least $\psi(y(q_+))$. Both aggregate bounds are simultaneously attainable by rows supported on $\{s, z\}$, making w transient. Consequently

$$g^{\mathrm{rob}}(P_0) = g^{\mathrm{rob}}(P_1) = \psi(y(q_+)), \tag{103}$$

and the gain is scalar from all three initial states.

Fix an arbitrary stationary policy π, write $\lambda := \pi(d \mid s)$, and let $\beta_u := \pi(u \mid z)$. Set

$$y_0(\beta) := \beta_a y(q_+) + \beta_b y(q_-) + (1 - \beta_a - \beta_b) y(q_1),$$
$$y_1(\beta) := \beta_a y(q_-) + \beta_b y(q_+) + (1 - \beta_a - \beta_b) y(q_1).$$

For an adversarial upper bound, use the two-state minimizing rows on $\{s, z\}$ for d and all z-actions, delete the navigation mass to w, and use the nominal rows at w. Starting at w gives only a transient prefix, and the unique recurrent class lies in $\{s, z\}$. Dropping the additional transitions to z caused by the other nondefault actions at s can only increase its gain. Therefore, uniformly in $x \in \{s, z, w\}$,

$$g^{\pi}_{\mathrm{rob},P_i}(x) \le \frac{\lambda y_i(\beta)}{\lambda\bar{x} + y_i(\beta)} \le \psi(y_i(\beta)). \tag{104}$$

Equation (100) gives $y_0(\beta) + y_1(\beta) \le y(q_+) + y(q_-)$. Concavity of ψ, exactly as in (90), now yields the statewise decision separation

$$\inf_{\pi,x}[\ell_0(\pi, x) + \ell_1(\pi, x)] \ge c_{\mathrm{dec}}\varepsilon. \tag{105}$$

Taking $y_i(\beta) \le y(q_+)$ in (104) also gives, for every i, π, x,

$$\begin{aligned}\ell_i(\pi, x) &\ge \psi(y(q_+)) - \frac{\lambda y(q_+)}{\lambda\bar{x} + y(q_+)} \\ &= \frac{(1-\lambda)y(q_+)^2}{[\bar{x} + y(q_+)][\lambda\bar{x} + y(q_+)]} \ge c_{\mathrm{rev}}(1-\lambda),\end{aligned} \tag{106}$$

where $c_{\mathrm{rev}} > 0$ is uniform for small ε.

Let $N_D(T)$ count actual nondefault actions at s, and let $N_z^{a,b}(T)$ count informative transitions from z. Conditional on the history, a nondefault action has probability $\mathbf{1}\{S_t = s\}(1-\lambda_t) \le 1-\lambda_t$; hence (106) gives

$$\mathbb{E}_i \operatorname{Reg}^{(i)}_{\mathrm{pol}}(T) \ge c_{\mathrm{rev}}\mathbb{E}_i N_D(T). \tag{107}$$

Under the nominal dynamics, every excursion from s is initiated by a nondefault action. At z every action returns to s with probability at least $r_{\min} := \min\{q_{\mathrm{l}}, q_0 - \varepsilon_0\} > 0$, and at w every action exits to $\{s, z\}$ with fixed positive probability. The geometric excursion argument therefore gives

$$\mathbb{E}_0 N_z^{a,b}(T) \le r_{\min}^{-1}\mathbb{E}_0 N_D(T).$$

Only the two hard rows differ, so the adaptive KL chain rule yields

$$\mathrm{KL}(\mathbb{P}_0^{\mathfrak{A},T} \| \mathbb{P}_1^{\mathfrak{A},T}) \le C_{\mathrm{kl}} r_{\min}^{-1}\varepsilon^2 \mathbb{E}_0 N_D(T). \tag{108}$$

With $\varepsilon = c_0 T^{-1/3}$, the same two-case argument as in Theorem 14 applies: either $\mathbb{E}_0 N_D(T) = \Omega(T^{2/3})$, in which case (107) proves the result, or the KL divergence in (108) is at most $1/2$, in which case Theorem 11 and (105) prove it.

Finally, every nonzero probability other than $q_\pm$ is fixed, and $q_\pm$ range over a compact interior interval. The strict inequalities above bound Δ_θ, $p_{\min,C}$ and Δ_{bell} away from zero. Fixed navigation and return probabilities bound D_0; compact fixed-face Bellman stability bounds H_{loc}. Taking B larger than these constants places both models in $\mathfrak{P}_f^{\mathrm{all}}(B)$ uniformly in T. The upper bound is Theorem 9. □

D.13 Information-theoretic lower limits

Proposition 13. *Let $f \in \mathcal{F} \setminus \mathcal{F}_{\mathrm{nc}}$ and choose $0 < q_{\mathrm{L}} < q_{\mathrm{H}} < 1$ such that*

$$\theta_f(q_{\mathrm{L}}) \le \delta < \theta_f(q_{\mathrm{H}}).$$

We may and do also require $q_{\mathrm{L}} + q_{\mathrm{H}} \le 1$: setting $q_* := \sup\{q : \theta_f(q) \le \delta\} < 1$, choose $q_{\mathrm{H}} \in (q_*, 1)$ and then $q_{\mathrm{L}} \in (0, 1 - q_{\mathrm{H}}]$ small enough that $\theta_f(q_{\mathrm{L}}) \le \delta$, which is possible because $\theta_f(0) = 0$ and θ_f is nondecreasing. Let $m := \underline{p}_f(q_{\mathrm{H}}) > 0$ and

$$\begin{aligned}\Delta_{\mathrm{del}} &:= \frac{2m}{(1+m)(2+m)}, \\ I_{\mathrm{del}} &:= \max\left\{\mathrm{kl}(q_{\mathrm{H}} \| q_{\mathrm{L}}), \mathrm{kl}(q_{\mathrm{L}} \| q_{\mathrm{H}})\right\}.\end{aligned}$$

There are two weakly communicating two-state kernels P_0, P_1 with $\delta < \delta_f^*(P_i)$ such that every history-dependent policy-valued learner satisfies

$$\max_{i\in\{0,1\}} \mathbb{E}_i \operatorname{Reg}_{\mathrm{pol}}^{(i)}(T) \geq \frac{T\Delta_{\mathrm{del}}}{2}\left(1-\sqrt{\frac{TI_{\mathrm{del}}}{2}}\right)_+.$$

Consequently, whenever $TI_{\mathrm{del}} \leq 1/2$, the lower bound is $T\Delta_{\mathrm{del}}/4$. This is a decision–information lower bound for resolving which action retains an undeletable transition. Two caveats are worth stating explicitly. First, q_{L} and q_{H} are fixed, so I_{del} is a positive constant of the instance and the displayed bound is informative only on the horizon range $T < 2/I_{\mathrm{del}}$; it is a finite-horizon separation, not a rate. Second, it does not assert a universal linear rate as the two rows approach the threshold, because m may simultaneously converge to zero. Within $\mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$ the asymptotic content is instead supplied by Proposition 9, whose two-point gap is shrunk with T.

Proof. Lemma 4 records that θ_f is nondecreasing on $[0,1]$. In particular,

$$\theta_f(1) \geq \theta_f(q_{\mathrm{H}}) > \delta.$$

Let $\mathcal{S} = \{s, z\}$ and $\mathcal{A} = \{0, 1\}$. Set

$$r(s,a) = 0, \qquad r(z,a) = 1 \qquad (a \in \mathcal{A}).$$

Both actions return deterministically from z to s:

$$P_i(s \mid z, a) = 1 \qquad (i \in \{0,1\},\ a \in \mathcal{A}).$$

At s, define

$$\begin{aligned} P_0(z \mid s, 0) &= q_{\mathrm{H}}, & P_0(z \mid s, 1) &= q_{\mathrm{L}}, \\ P_1(z \mid s, 0) &= q_{\mathrm{L}}, & P_1(z \mid s, 1) &= q_{\mathrm{H}}. \end{aligned}$$

The complementary probabilities return to s.

Each nominal graph is communicating: choosing at s the action with probability $q_{\mathrm{H}} > 0$ reaches z, and every action at z returns to s. The edge from s to z under the high action belongs to the robust core because $\theta_f(q_{\mathrm{H}}) > \delta$. The deterministic edge from z to s also belongs to the robust core because $\theta_f(1) > \delta$. The remaining row is nonempty as well: since $q_{\mathrm{L}} + q_{\mathrm{H}} \leq 1$ we have $1 - q_{\mathrm{L}} \geq q_{\mathrm{H}}$, so monotonicity of θ_f gives $\theta_f(1 - q_{\mathrm{L}}) \geq \theta_f(q_{\mathrm{H}}) > \delta$ and the self-loop at s under the low action survives. Thus, in either model, the robust core contains $s \to z$ under the high action, $s \to s$ under the low action, and $z \to s$ under every action. Its only nonempty closed set is $\mathcal{S}$. Hence $\mathcal{G}_\wedge(P_i)$ is weakly communicating. Theorem 1 now gives

$$\delta < \delta_f^*(P_i), \qquad i \in \{0,1\}.$$

By the exact deletion formula,

$$\underline{p}_f(q_{\mathrm{L}}) = 0, \qquad \underline{p}_f(q_{\mathrm{H}}) = m > 0.$$

Consider a stationary randomized policy that selects action 0 at s with probability $\alpha \in [0,1]$. Rectangularity lets the adversary minimize the probability of entering the rewarding state separately in the two action rows. Therefore the smallest transition probability from s to z is

$$x_0(\alpha) = \alpha m \quad \text{under } P_0, \qquad x_1(\alpha) = (1-\alpha)m \quad \text{under } P_1.$$

At z, the nominal row returning to s is feasible. Any created self-loop at z can only increase the stationary fraction of time in the rewarding state, so the minimizing kernel returns to s with probability one. A two-state chain with transition probability x from s to z and probability one from z to s has stationary distribution

$$\mu_x(s) = \frac{1}{1+x}, \qquad \mu_x(z) = \frac{x}{1+x}.$$

Consequently, with $\phi(x) := x/(1+x)$,

$$g_{\mathrm{rob},P_0}^{\pi_\alpha} = \phi(\alpha m), \qquad g_{\mathrm{rob},P_1}^{\pi_\alpha} = \phi((1-\alpha)m),$$

and

$$g^{\mathrm{rob}}(P_0) = g^{\mathrm{rob}}(P_1) = \phi(m).$$

All these gains are scalar: if the minimizing probability from s to z is positive the chain is irreducible, while if it is zero then s is the unique recurrent state and z returns to s.

The function ϕ is increasing and concave. Hence

$$\begin{aligned}\Delta_{\mathrm{dec}}(P_0, P_1) &= \min_{\alpha\in[0,1]} \{2\phi(m) - \phi(\alpha m) - \phi((1-\alpha)m)\} \\ &= 2\phi(m) - 2\phi(m/2) \\ &= \frac{2m}{(1+m)(2+m)} = \Delta_{\mathrm{del}}.\end{aligned}$$

The second equality follows because concavity implies

$$\phi(\alpha m) + \phi((1-\alpha)m) \le 2\phi(m/2),$$

with equality at $\alpha = 1/2$.

It remains to bound the information. By the chain rule for relative entropy, the KL divergence between the two interaction laws equals the sum of expected conditional row divergences. The rows at z agree exactly. At s, the two Bernoulli parameters are $q_{\mathrm{H}}, q_{\mathrm{L}}$ in one order or the other. Therefore every one-step conditional divergence is at most I_{del}, and

$$\mathrm{KL}\Big(\mathbb{P}_0^{\mathfrak{A},T} \Big\| \mathbb{P}_1^{\mathfrak{A},T}\Big) \le T I_{\mathrm{del}}.$$

Substitution into Theorem 10 proves the stated lower bound. If $TI_{\mathrm{del}} \le 1/2$, the parenthetical factor is at least $1/2$, giving $T\Delta_{\mathrm{del}}/4$. □

Proposition 14. *Let $f \in \mathcal{F} \setminus \mathcal{F}_{\mathrm{ne}}$ and $0 < \delta < \theta_f(1)$. There exist two-state kernels P_0, P_1, a common reward, a common initial state s_0, and $\Delta_{\mathrm{hid}} > 0$ such that, for interactions starting from s_0,*

$$\begin{aligned}&\mathcal{G}(P_i) \in \mathcal{M}, \quad C(P_i) \neq \mathcal{S}, \quad \delta < \delta_f^*(P_i) \le \delta_f^\dagger(P_i), \\ &\mathbb{P}_{P_0}^{\mathfrak{A},T} = \mathbb{P}_{P_1}^{\mathfrak{A},T} \qquad \forall\, \mathfrak{A},\ \forall\, T, \\ &\inf_{\mathfrak{A}\in\mathfrak{A}_{\mathrm{pol}}} \max_{i\in\{0,1\}} \mathbb{E}_{P_i}^{\mathfrak{A}} \operatorname{Reg}_{\mathrm{pol}}^{(i)}(T) \ \ge\ \tfrac{1}{2}\Delta_{\mathrm{hid}} T.\end{aligned}$$

The nominal laws are identical while the robustly optimal actions differ. For $f \in \mathcal{F}_{\mathrm{nc}} \setminus \mathcal{F}_{\mathrm{ne}}$ the premise $\delta < \theta_f(1)$ holds at every finite radius.

Proof. We first construct a created transition with positive mass. Since $f \notin \mathcal{F}_{\mathrm{ne}}$, the recession slope $f'(\infty)$ is finite. For $x \in [0,1]$, the cost of changing the nominal row $(1,0)$ into $(1-x, x)$ is

$$d_{\mathrm{cr}}(x) := D_f\big((1-x,x)\|(1,0)\big) = f(1-x) + x f'(\infty).$$

Because f is continuous at 1 and $f(1) = 0$,

$$\lim_{x\downarrow 0} d_{\mathrm{cr}}(x) = 0.$$

Since $\delta > 0$, some $x > 0$ satisfies $d_{\mathrm{cr}}(x) \le \delta$. Compactness and lower semicontinuity ensure that the maximum

$$\overline{x} := \max\{x \in [0,1] : d_{\mathrm{cr}}(x) \le \delta\}$$

exists, and the displayed limit together with $\delta > 0$ gives $\overline{x} > 0$.

We next construct two protected exit rows with different robust minima. Set $y(q) := \underline{p}_f(q)$. The inequality $\delta < \theta_f(1)$ implies

$$y(1) > 0.$$

Indeed, if $y(1) = 0$, compactness would provide feasible $p_n \downarrow 0$; lower semicontinuity would then imply $\theta_f(1) \le \delta$, a contradiction. Let $q_{\mathrm{H}} := 1$ and $y_{\mathrm{H}} := y(1)$.

Suppose first that $f \in \mathcal{F}_{\mathrm{nc}}$. Then every positive nominal mass is undeletable, so $y(q) > 0$ for every $q > 0$. The nominal row itself is feasible, hence $y(q) \le q$. Choose $q_{\mathrm{L}} \in (0, y_{\mathrm{H}})$. Then

$$0 < y_{\mathrm{L}} := y(q_{\mathrm{L}}) \le q_{\mathrm{L}} < y_{\mathrm{H}}.$$

Now suppose $f \notin \mathcal{F}_{\mathrm{nc}}$. Because also $f \notin \mathcal{F}_{\mathrm{ne}}$, both $f(0)$ and $f'(\infty)$ are finite. Thus θ_f is continuous on $[0,1]$. Lemma 4 shows that it is nondecreasing. Since

$$\theta_f(0) = 0 < \delta < \theta_f(1),$$

Let $q_\star := \sup\{q : \theta_f(q) \le \delta\} \in (0,1)$. By continuity $\theta_f(q_\star) = \delta$, and $\theta_f(q) > \delta$ for every $q > q_\star$. Choose $q_n \downarrow q_\star$ with $\theta_f(q_n) > \delta$. Define

$$\lambda_n := 1 - \frac{\delta}{\theta_f(q_n)}, \qquad p_n := \lambda_n q_n.$$

The distribution $(p_n, 1-p_n)$ is the convex combination

$$(p_n, 1-p_n) = (1-\lambda_n)(0,1) + \lambda_n(q_n, 1-q_n).$$

Convexity of the f-divergence in its first argument gives

$$\begin{aligned} D_f\big((p_n,1-p_n)\|(q_n,1-q_n)\big) &\le (1-\lambda_n)\theta_f(q_n) + \lambda_n \cdot 0 \\ &= \delta. \end{aligned}$$

Therefore $y(q_n) \le p_n$. Since $\theta_f(q_n) \to \delta$, we have $\lambda_n \to 0$, and hence $p_n \to 0$. On the other hand, $\theta_f(q_n) > \delta$ implies $y(q_n) > 0$. Choose n large enough that $y(q_n) < y_{\mathrm{H}}$, and set

$$q_{\mathrm{L}} := q_n, \qquad y_{\mathrm{L}} := y(q_n).$$

In both cases we have obtained

$$0 < y_{\mathrm{L}} < y_{\mathrm{H}}, \qquad \theta_f(q_{\mathrm{L}}) > \delta, \qquad \theta_f(q_{\mathrm{H}}) > \delta. \tag{109}$$

We now define the indistinguishable nominal kernels. Let $\mathcal{S} = \{s, z\}$, $\mathcal{A} = \{0,1\}$, and

$$r(s,a) = 1, \qquad r(z,a) = 0 \qquad (a \in \mathcal{A}).$$

Both actions self-loop at s under both nominal kernels:

$$P_i(s \mid s,a) = 1 \qquad (i \in \{0,1\},\ a \in \mathcal{A}).$$

At z, let the probability of exiting to s be

$$\begin{aligned} P_0(s \mid z,0) &= q_{\mathrm{H}}, & P_0(s \mid z,1) &= q_{\mathrm{L}}, \\ P_1(s \mid z,0) &= q_{\mathrm{L}}, & P_1(s \mid z,1) &= q_{\mathrm{H}}. \end{aligned}$$

The complementary probabilities self-loop at z.

Under either nominal kernel, s is the unique recurrent state. From z, every deterministic policy exits to s with probability at least $q_{\mathrm{L}} > 0$, after which the process never returns. Thus z is transient under every deterministic policy, and each nominal MDP is weakly communicating with $C(P_i) = \{s\}$.

The edge $s \to s$ has nominal mass one and survives because $\theta_f(1) > \delta$. Every action at z has an edge to s whose mass is either q_{L} or q_{H}, and both survive by inequality (109). Hence $\mathcal{G}_\wedge(P_i)$ contains

$$s \longrightarrow s, \qquad z \longrightarrow s \quad \text{under every action.}$$

This core graph is weakly communicating with recurrent set $\{s\}$. Theorem 1 therefore gives

$$\delta < \delta_f^*(P_i).$$

Proposition 2 then gives $\delta_f^*(P_i) \le \delta_f^\dagger(P_i)$.

Set $s_{\mathrm{init}} := s$. Starting from this state, the nominal state remains s forever under every algorithm and under both kernels. Rewards are also identical. Thus

$$\mathbb{P}_0^{\mathfrak{A},T} = \mathbb{P}_1^{\mathfrak{A},T}, \qquad \mathrm{KL}\Big(\mathbb{P}_0^{\mathfrak{A},T} \Big\| \mathbb{P}_1^{\mathfrak{A},T}\Big) = 0.$$

Finally, we compute the robust values, decision separation, and coverage. At s, the adversary minimizes reward by maximizing the created probability of moving to z, so it chooses $\overline{x}$. At z, it

minimizes reward by minimizing the exit probability to the rewarding state s. If a stationary policy chooses action 0 at z with probability α, rectangularity gives

$$y_0(\alpha) = \alpha y_{\mathrm{H}} + (1-\alpha) y_{\mathrm{L}},$$
$$y_1(\alpha) = \alpha y_{\mathrm{L}} + (1-\alpha) y_{\mathrm{H}}.$$

A two-state chain with transition probability $\overline{x}$ from s to z and transition probability y from z to s has stationary mass

$$\mu_y(s) = \frac{y}{\overline{x}+y}, \qquad \mu_y(z) = \frac{\overline{x}}{\overline{x}+y}.$$

Writing $\psi(y) := y/(\overline{x}+y)$, we obtain

$$g^{\pi_\alpha}_{\mathrm{rob},P_0} = \psi(y_0(\alpha)), \qquad g^{\pi_\alpha}_{\mathrm{rob},P_1} = \psi(y_1(\alpha)),$$

and

$$g^{\mathrm{rob}}(P_0) = g^{\mathrm{rob}}(P_1) = \psi(y_{\mathrm{H}}).$$

Because $\overline{x} > 0$ and $y_{\mathrm{L}} > 0$, every minimizing chain is irreducible, so all gains are scalar.

The function ψ is strictly increasing and concave. Since $y_0(\alpha) + y_1(\alpha) = y_{\mathrm{H}} + y_{\mathrm{L}}$, concavity gives

$$\psi(y_0(\alpha)) + \psi(y_1(\alpha)) \le 2\psi\left(\frac{y_{\mathrm{H}} + y_{\mathrm{L}}}{2}\right).$$

Consequently,

$$\Delta_{\mathrm{dec}}(P_0, P_1) = 2\psi(y_{\mathrm{H}}) - 2\psi\left(\frac{y_{\mathrm{H}} + y_{\mathrm{L}}}{2}\right)$$
$$=: \Delta_{\mathrm{hid}} > 0.$$

Applying Theorem 10 with zero KL divergence gives

$$\max_i \mathbb{E}_i \operatorname{Reg}^{(i)}_{\mathrm{pol}}(T) \ge \frac{\Delta_{\mathrm{hid}}}{2} T.$$

Finally, under P_i the stationary mass of z is zero. The ambiguity kernel used above has stationary mass

$$\mu^{\pi}_{\widetilde{Q}}(z) = \frac{\overline{x}}{\overline{x}+y} > 0.$$

Thus the ratio in the definition of $c_{\mathrm{cov}}(P_i)$ contains a strictly positive numerator and a zero denominator. Hence

$$c_{\mathrm{cov}}(P_i) = +\infty.$$

This completes the construction and the proof. □

Proposition 15.

$$\delta \ge \delta^{\dagger}_f(P) \implies (\exists \widetilde{Q} \in \mathcal{U}^P(\delta))(\exists r \in [0,1]^{\mathcal{S}\times\mathcal{A}}) : \mathcal{G}(\widetilde{Q}) \notin \mathcal{M} \text{ and } g^*_{\widetilde{Q}} \notin \mathbb{R}\mathbf{1}.$$

Thus $\delta < \delta^{\dagger}_f(P)$ holds exactly when every ambiguity kernel has a state-independent nominal optimal gain for every bounded reward. Proposition 15 does not assert failure of the robust Bellman equation at or above the threshold.

Proof. Theorem 2 gives $\widetilde{Q} \in \mathcal{U}^P(\delta)$ whose support graph is not weakly communicating. By the graph definition, there are disjoint nonempty sets $I, J \subseteq \mathcal{S}$ such that I is closed under every action and J is closable. Thus there is a deterministic policy π_J satisfying

$$\mathcal{G}(\widetilde{Q})(s, \pi_J(s)) \subseteq J, \qquad s \in J.$$

Define

$$r(s,a) := \mathbf{1}\{s \in J\}.$$

Fix $i \in I$. Since I is closed under every action, a trajectory starting from i remains in I under every policy. Because $I \cap J = \varnothing$, every reward on that trajectory is zero, and

$$g^*_{\widetilde{Q}}(i) = 0.$$

Fix $j \in J$. Under π_J, a trajectory starting from j remains in J and receives reward one at every time. Therefore

$$g^{\pi_J}_{\widetilde{Q}}(j) = 1.$$

Rewards are bounded above by one, so

$$g^*_{\widetilde{Q}}(j) = 1.$$

Hence the optimal gain takes different values at i and j. This proves that no conclusion asserting a scalar optimal gain for every ambiguity kernel can extend beyond $\delta^\dagger_f(P)$. □

D.14 Main-text component summaries

Theorem 4. Let $m_T^{\mathrm{BC}} := \lceil \sqrt{T} \rceil$, $m_T := \lceil T^{2/3} \rceil$, and $\eta_T := \beta_{\alpha/3}(n_T(P, \alpha))$, where $n_T(P, \alpha)$ is the uniform-exploration sample floor. The finite readiness thresholds are given in Appendix D. For $f \in \mathcal{F}_{\mathrm{ne}}$, let $\mathfrak{C}_{\mathrm{ne}}(P_R) := H_{\mathrm{loc}}(P_R) K_{\mathrm{face}}(P_R) C_{\mathrm{occ}}(P_R)$.

(i) If $f \in \mathcal{F}_{\mathrm{ne}}$ and $\delta < \delta^{\mathrm{bell}}_f(P)$, then after Bellman-completion readiness, Algorithm 3 satisfies

$$\begin{aligned}\mathbb{E}\,\mathrm{Reg}_{\mathrm{pol}}(T) \le m_T^{\mathrm{BC}} &+ \widetilde{O}\Big(\mathfrak{C}_{\mathrm{ne}}(P_R)\sqrt{d_R A T[S + \log(SAT/\alpha)]}\Big) \\ &+ \widetilde{O}\big(\mathfrak{C}_{\mathrm{ne}}(P_R) M_{\mathrm{mix}}(P_R) d_R^2 A\big) + \alpha T.\end{aligned}$$

No assumption $C(P) = \mathcal{S}$ is made. Moreover $C_{\mathrm{occ}} \le 1 + D_{\mathrm{cmp}}$, $M_{\mathrm{mix}} \le 2D_{\mathrm{cmp}}$, and $K_{\mathrm{face}} \le 2/p_{\min,R}$, so all displayed constants are controlled by the primitive family parameters.

(ii) If $\mathcal{G}(P) \in \mathcal{M}$, $\delta < \delta^{\mathrm{bell}}_f(P)$, and either $f \in \mathcal{F}_{\mathrm{ne}}$ or $C(P) = \mathcal{S}$, then, for horizons above the finite threshold specified in Theorem 9, Algorithm 1 satisfies with probability at least $1 - \alpha$

$$\mathrm{Reg}_{\mathrm{pol}}(T) \le m_T + 2(T - m_T) H_{\mathrm{loc}}(P_R) \varpi^{P_R}_{f,\delta}(\eta_T).$$

For each fixed instance this bound is $o(T)$. Under the additional support and accuracy conditions of Theorem 9, the same probability guarantee gives

$$\widetilde{O}\Big([1 + H_{\mathrm{loc}}(P_R) K_{\mathrm{face}}(P_R) \sqrt{D_0(P) d_R A[\mathcal{S} + \log(SAT/\alpha)]}] T^{2/3}\Big).$$

All bounds hold uniformly in the initial state. For each fixed instance, taking $\alpha = T^{-2}$ gives expected regret $\widetilde{O}(\sqrt{T})$ in (i) and $\widetilde{O}(T^{2/3})$ in (ii).

Proof. Part (i) is Theorem 13, followed by Lemma 10 and the fixed-face inequality $K_{\mathrm{face}} \le 2/p_{\min,R}$. Part (ii) is Theorem 9; its proof establishes the displayed gap for every $x \in R(P)$, so summing the per-step comparator preserves the bound and the commit step is uniform in the state. □

Proposition 16. *For each nonempty case below, there exists a family with fixed finite state and action spaces whose minimax expected policy regret satisfies the stated lower bound. The family parameters, including positive lower bounds on the structural margins, are independent of T.*

Generator and regime	*Additional condition*	*Lower bound*
$f \in \mathcal{F}_{\mathrm{ne}}$, *common core*	*none*	$\Omega(\sqrt{T})$
$f \in \mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$, all *regime*	$\delta^*_f \le \delta < \delta^\dagger_f$	$\Omega(\sqrt{T})$
$f \in \mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$, bell *regime*	$\delta^\dagger_f \le \delta < \delta^{\mathrm{bell}}_f$	$\Omega(\sqrt{T})$
$f \notin \mathcal{F}_{\mathrm{ne}}$, *common core*	$C(P) = \mathcal{S}$, $\delta < \theta_f(1)$	$\Omega(T^{2/3})$
$f \notin \mathcal{F}_{\mathrm{ne}} \cup \mathcal{F}_{\mathrm{nc}}$, all *regime*	$C(P) = \mathcal{S}$, $\delta < \theta_f(1)$	$\Omega(T^{2/3})$
$f \notin \mathcal{F}_{\mathrm{ne}} \cup \mathcal{F}_{\mathrm{nc}}$, bell *regime*	$C(P) = \mathcal{S}$, $\delta < \theta_f(1)$	$\Omega(T^{2/3})$
$f \notin \mathcal{F}_{\mathrm{ne}}$, *hidden recurrence*	$C(P) \ne \mathcal{S}$, $\delta < \theta_f(1)$	$\Omega(T)$

For $S \geq 3$ and $A \geq 2$, the no-creation common-core construction gives a regret lower bound of $\Omega_{f,\delta}(\sqrt{\lfloor S/3 \rfloor AT})$ and satisfies $D_{\mathrm{cmp}}(P_R) = \Theta(S)$.

Proof. The three no-creation rows are Propositions 9 and 10. The three communicating creation rows are Theorem 14 and Propositions 12 and 11. The hidden-recurrence row is Proposition 14. Each cited construction states the strict radius inequalities and uniform bounds on its family constants; the parallelized refinement and $D_{\mathrm{cmp}} = \Theta(S)$ are part of Proposition 9. □

Proposition 1. Suppose $f \in \mathcal{F}_{\mathrm{ne}}$, $S \geq 3$, and $A \geq 2$. For $B \geq C_{f,\delta} SA$ and all sufficiently large T,

$$c_{f,\delta} \sqrt{\lfloor S/3 \rfloor AT} \leq \mathfrak{R}_f^{\mathrm{core}}(T; B) \leq \widetilde{O}_{f,\delta}\Big(B^3 \sqrt{SAT[S + \log(SAT)]} + B^4 S^2 A \Big).$$

The bounds match in their exponent of T, while a polynomial gap remains in their dependence on S, A, and B.

Separately, the navigation quantities satisfy

$$D_{\mathrm{cmp}}(P_R) \leq (d_R - 1) p_{\min,R}^{-(d_R-1)},$$

and for every $L \geq 1$ and $p \in (0, 1/2]$ there is a communicating one-action chain on $L+1$ states satisfying

$$p_{\min} = p, \qquad D_{\mathrm{cmp}}(P) \geq \mathbb{E}_0 \tau_L = \frac{1 - p^L}{(1-p)p^L}.$$

Hence the exponent of $p_{\min}$ is unavoidable for a worst-case conversion to hitting time. This is a navigation lower bound, not a regret lower bound; the regret theorem keeps D_{cmp} primitive.

Proof. The regret sandwich is Theorem 13 combined with the parallelized part of Proposition 9 and the family bounds on $H_{\mathrm{loc}}, K_{\mathrm{face}}, C_{\mathrm{occ}}$, and M_{mix}. □

D.15 Assembly of the complete phase diagram

Lemma 20 (Exact full-simplex saturation)**.** *Suppose $\delta \geq \theta_f(1)$. Then, for every nominal row q, policy π, kernel P, and initial state s,*

$$\begin{aligned} \mathcal{U}^q(\delta) &= \Delta(\mathcal{S}), & \sigma_\delta^f(q,u) &= \min_x u(x), \\ g_{\mathrm{rob},P}^\pi(s) &= \min_x \sum_a \pi(a \mid x) r(x,a), & g^{\mathrm{rob}}(P) &= \min_x \max_a r(x,a). \end{aligned} \tag{110}$$

Consequently any selector $\pi^r(x) \in \arg\max_a r(x,a)$ has $\mathrm{Reg}_{\mathrm{pol}}(T) = 0$ for every T. If $S \geq 2$, then also $\delta_f^*(P) \leq \delta_f^\dagger(P) \leq \theta_f(1)$ and $\delta_f^{\mathrm{bell}}(P) = +\infty$.

Proof. Fix a state j. If $q_j > 0$, then

$$D_f(\mathbf{e}_j \| q) = \theta_f(1 - q_j) \leq \theta_f(1) \leq \delta,$$

whereas if $q_j = 0$, the boundary conventions give $D_f(\mathbf{e}_j \| q) = f(0) + f'(\infty) = \theta_f(1) \leq \delta$. Thus the convex set $\mathcal{U}^q(\delta)$ contains every vertex of the simplex and hence equals $\Delta(\mathcal{S})$, proving the first two identities in (110). When $S \geq 2$, choose distinct states i, j. At radius $\theta_f(1)$, rectangularity permits a kernel whose every action at i stays at i and which has an action at j that stays at j. Its graph is not weakly communicating, so Theorem 2 and Proposition 2 give $\delta_f^*(P) \leq \delta_f^\dagger(P) \leq \theta_f(1)$. Finite $\theta_f(1)$ makes creation possible, hence $\delta_f^{\mathrm{bell}}(P) = +\infty$ by Theorem 3. Let $r_\pi(x) := \sum_a \pi(a \mid x) r(x,a)$ and choose $x_\pi \in \arg\min_x r_\pi(x)$. Every reward stream under π is bounded below in expectation by $r_\pi(x_\pi)$, while the feasible rectangular kernel whose every row is $\mathbf{e}_{x_\pi}$ moves to x_π in one step and remains there. Therefore $g_{\mathrm{rob},P}^\pi(s) = \min_x r_\pi(x)$ for every s. Maximizing separately over the action distribution at each state gives $g^{\mathrm{rob}}(P) = \min_x \max_a r(x,a)$, attained by π^r. Every summand in (5) then vanishes. □

Theorem 5. Fix $f \in \mathcal{F}$ and $\delta > 0$.

(i) *All scalar nonsaturated cells.* Suppose $\delta < \theta_f(1)$. For every nonempty regime indexed by $\xi \in \Xi_f$, there are fixed finite witness dimensions S_ξ, A_ξ and constants $B_{0,\xi}, c_{f,\delta,\xi} > 0$ such that, for every $B \geq B_{0,\xi}$, there is a finite $T_{0,\xi}(B, f, \delta)$ for which

$$c_{f,\delta,\xi} T^{\alpha_f} \;\leq\; \mathfrak{R}_f^{\xi}(T; B) \;\leq\; \widetilde{O}_{B,f,\delta,S_\xi,A_\xi}\big(T^{\alpha_f}\big), \qquad T \geq T_{0,\xi}. \tag{111}$$

At the fixed finite witness dimensions in the preceding display,

$$\mathfrak{R}_f^{\xi}(T; B) = \begin{cases} \widetilde{\Theta}(\sqrt{T}), & f \in \mathcal{F}_{\mathrm{ne}}, \\ \widetilde{\Theta}(T^{2/3}), & f \notin \mathcal{F}_{\mathrm{ne}}. \end{cases}$$

The first line imposes no assumption $C(P) = \mathcal{S}$: Algorithm 3 learns on P_R. Thus the neither-operation class has exponent $1/2$ at every finite radius, and the deletion-only class has exponent $1/2$ throughout all three cells

$$\delta < \delta_f^*(P), \qquad \delta_f^*(P) \leq \delta < \delta_f^\dagger(P), \qquad \delta_f^\dagger(P) \leq \delta < \delta_f^{\mathrm{bell}}(P).$$

With creation and $C(P) = \mathcal{S}$, the exponent is $2/3$ in every nonempty cell below $\theta_f(1)$. Hence δ_f^* and $\delta_f^\dagger$ are exact structural transitions but not horizon-exponent transitions.

(ii) *Creation without nominal communication.* If $f \notin \mathcal{F}_{\mathrm{ne}}$ and $\delta < \theta_f(1)$, there exist two weakly communicating kernels P_0, P_1 with $C(P_i) \neq \mathcal{S}$, a common reward r, and a common initial state s_0 such that, for some $c_{f,\delta} > 0$ and all sufficiently large T,

$$c_{f,\delta} T \leq \inf_{\mathfrak{A} \in \mathfrak{A}_{\mathrm{pol}}} \max_{i \in \{0,1\}} \mathbb{E}^{\mathfrak{A}}_{P_i, r, s_0} \operatorname{Reg}_{\mathrm{pol}}(T) \leq T.$$

Thus nominal weak communication alone does not guarantee uniform sublinear regret when creation is possible and the ambiguity sets are not saturated.

(iii) *Full-simplex saturation.* If $\delta \geq \theta_f(1)$, the identities in (110) hold and a policy choosing from $\arg\max_a r(s, a)$ at every state has zero regret for every P, so the minimax risk over any nonempty class of nominal kernels is zero. This finite-radius case occurs only in the both-operations class.

(iv) *Failure of the scalar comparator.* If $f \in \mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$ and $\delta \geq \delta_f^{\mathrm{bell}}(P)$, then some bounded reward has state-dependent robust optimal gain. For such a reward, no common scalar comparator exists, so the regret in (5) is not defined.

Proof. For (i), Theorem 13 gives the no-creation upper bound on P_R in every scalar cell; Propositions 9 and 10 give the core and two annular lower bounds. For creation, Theorem 9 gives every communicating upper bound, while Theorem 14 and Propositions 12 and 11 give the three matching lower bounds. Taking the largest finite supporting burn-in proves (7). Part (ii) is Proposition 14 together with the unit per-step upper bound. Part (iii) is Lemma 20, and part (iv) is Theorem 3. □

D.16 Divergence-class specialization

Proposition 17. *Suppose $f \in \mathcal{F}_{\mathrm{nc}} \cap \mathcal{F}_{\mathrm{ne}}$, $\mathcal{G}(P) \in \mathcal{M}$, and $\delta < +\infty$. Then*

$$\mathcal{G}_\wedge(P) = \mathcal{G}(\widetilde{Q}) = \mathcal{G}(P), \qquad \delta_f^*(P) = \delta_f^\dagger(P) = \delta_f^{\mathrm{bell}}(P) = +\infty,$$

for every $\widetilde{Q} \in \mathcal{U}^P$. Theorem 9 therefore applies on $R(P) = C(P)$ without any additional structural, radius, or coverage condition and gives $\operatorname{Reg}_{\mathrm{pol}}(T) = o(T)$. More sharply, Theorem 13 and Proposition 9 give the minimax exponent $1/2$ at every finite radius on parameter-bounded families, with no assumption $C(P) = \mathcal{S}$. Neither the deletion-testing lower bound nor either creation lower bound applies in this class.

Proof. Because $f(0) = f'(\infty) = +\infty$, no nominal edge can be deleted and no nonnominal edge can be created. Thus

$$\mathcal{G}_\wedge(P) = \mathcal{G}(\widetilde{Q}) = \mathcal{G}(P) \qquad (\widetilde{Q} \in \mathcal{U}^P).$$

Every nominal edge has infinite deletion cost. Taking $H = \mathcal{G}(P) \in \mathcal{M}$ in the definition of $\delta_f^*(P)$ gives $\delta_f^*(P) = +\infty$. Every ambiguity kernel has the same weakly communicating graph, so Theorem 2 gives $\delta_f^\dagger(P) = +\infty$. Equation (4) also gives $\delta_f^{\mathrm{bell}}(P) = +\infty$ because the fixed weakly communicating graph admits no disjoint closed–closable obstruction. Since $f \in \mathcal{F}_{\mathrm{ne}}$, the relevant set in Theorem 9 is $R(P) = C(P)$. Its Bellman condition holds at every finite radius because $\delta_f^{\mathrm{bell}}(P) = +\infty$. Theorem 9 therefore gives $o(T)$ regret, and the final display of Theorem 9 gives the explicit fixed-face rate once its support and neighborhood thresholds hold. Since $\delta_f^{\mathrm{bell}}(P) = +\infty$, Theorem 13 applies to the recurrent restriction P_R at every finite radius and gives $\widetilde{O}(\sqrt{T})$ without requiring $C(P) = \mathcal{S}$. Proposition 9 gives the matching $\Omega(\sqrt{T})$ lower bound on a communicating subfamily, hence on the larger family as well. The premises of the deletion and creation lower bounds fail respectively because $f \in \mathcal{F}_{\mathrm{nc}}$ and $f \in \mathcal{F}_{\mathrm{ne}}$. □

Proposition 18. *Suppose $f \in \mathcal{F}_{\mathrm{ne}} \setminus \mathcal{F}_{\mathrm{nc}}$ and $\mathcal{G}(P) \in \mathcal{M}$. No edge outside $\mathcal{G}(P)$ can be created. The robust Bellman equation admits a scalar additive eigenpair for every bounded reward exactly when*

$$\delta < \delta_f^{\mathrm{bell}}(P).$$

Under this condition Theorem 9 applies on $R(P) = C(P)$ and gives $\mathrm{Reg}_{\mathrm{pol}}(T) = o(T)$, including throughout $\delta_f^*(P) \le \delta < \delta_f^{\mathrm{bell}}(P)$. No coverage condition is needed. The deletion-testing term in Proposition 13 is unavoidable. Theorem 13, Proposition 9, and Proposition 10 give the matched minimax exponent $1/2$ throughout the entire scalar regime $\delta < \delta_f^{\mathrm{bell}}(P)$, with no assumption $C(P) = \mathcal{S}$, separately in the core and on both sides of $\delta_f^\dagger(P)$. Within the second annulus, $\delta \ge \delta_f^\dagger(P)$ permits individual ambiguity kernels with state-dependent nominal optimal gain, but does not obstruct the robust Bellman comparator or the square-root completion. If $\delta \ge \delta_f^{\mathrm{bell}}(P)$, the failure reward in Theorem 3 makes the robust optimal gain itself state dependent and precludes a scalar robust eigenpair.

Proof. Since $f \in \mathcal{F}_{\mathrm{ne}}$, Theorem 1 gives

$$\mathcal{G}(\widetilde{Q}) \subseteq \mathcal{G}(P) \qquad (\widetilde{Q} \in \mathcal{U}^P).$$

Thus all nominal and ambiguity dynamics based in $C(P)$ stay in $C(P)$. Under $\delta < \delta_f^{\mathrm{bell}}(P)$, Lemma 18 proves $\delta < \delta_f^{\mathrm{bell}}(P_R)$ for the restricted kernel. Uniform exploration identifies $C(P)$ and estimates only its rows. Theorem 9 consequently gives $o(T)$ regret without a common robust core and without a coverage comparison. The proof of Theorem 9 does not invoke $\delta_f^*(P)$, so the conclusion applies throughout $\delta_f^*(P) \le \delta < \delta_f^{\mathrm{bell}}(P)$.

Throughout the scalar regime $\delta < \delta_f^{\mathrm{bell}}(P)$, Proposition 8 operates on the recurrent restriction and Theorem 13 gives the $\widetilde{O}(\sqrt{T})$ upper bound. Propositions 9 and 10 give matching $\Omega(\sqrt{T})$ lower bounds in the core and separately on both sides of $\delta_f^\dagger(P)$.

Because $f \notin \mathcal{F}_{\mathrm{nc}}$, Proposition 13 constructs hard instances inside this class. If instead $\delta \ge \delta_f^{\mathrm{bell}}(P)$, Theorem 3 supplies a bounded reward for which the robust optimal gain is state dependent. This proves both claimed limits. □

Proposition 19. *Suppose $f \in \mathcal{F}_{\mathrm{nc}} \setminus \mathcal{F}_{\mathrm{ne}}$, $\mathcal{G}(P) \in \mathcal{M}$, and $\delta < +\infty$. Then*

$$\mathcal{G}_\wedge(P) = \mathcal{G}(P) \subseteq \mathcal{G}(\widetilde{Q}), \qquad \delta_f^*(P) = \delta_f^\dagger(P) = \delta_f^{\mathrm{bell}}(P) = +\infty,$$

for every $\widetilde{Q} \in \mathcal{U}^P$. If the nominal kernel is communicating, that is, $C(P) = \mathcal{S}$, Theorem 9 applies on all of $\mathcal{S}$ at every finite radius and gives $\mathrm{Reg}_{\mathrm{pol}}(T) = o(T)$ without a coverage condition. This nominal communication requirement is minimax-sharp: if it is dropped, Proposition 14 gives weakly communicating instances with $C(P) \neq \mathcal{S}$ whose nominal laws are indistinguishable but whose robustly optimal actions disagree, forcing $\Omega(T)$ regret. Within the communicating regime, Theorem 14 supplies a two-state subproblem with minimax rate $\widetilde{\Theta}(T^{2/3})$, even under $\delta < \delta_f^*(P)$.

Proof. Because $f(0) = +\infty$, every positive nominal edge is undeletable, so

$$\mathcal{G}_\wedge(P) = \mathcal{G}(P).$$

Every ambiguity support contains this graph. Since $\mathcal{G}(P)$ is weakly communicating, every such supergraph is weakly communicating. All nominal edges have infinite deletion cost, and therefore

$$\delta_f^*(P) = \delta_f^\dagger(P) = +\infty.$$

Because $f \notin \mathcal{F}_{\mathrm{ne}}$, Theorem 3 also gives $\delta_f^{\mathrm{bell}}(P) = +\infty$. If $C(P) = \mathcal{S}$, the relevant set in Theorem 9 is all of $\mathcal{S}$, and every row is sampled by uniform exploration. Theorem 9 therefore gives $o(T)$ regret at every finite radius, without requiring $c_{\mathrm{cov}}(P) < +\infty$.

Because $f \notin \mathcal{F}_{\mathrm{ne}}$ and $\theta_f(1) = +\infty$, every premise of Proposition 14 holds. Proposition 14 constructs two weakly communicating instances in this class with proper recurrent set $C(P_i) = \{s\} \neq \mathcal{S}$ and proves an $\Omega(T)$ lower bound for every history-dependent policy-valued learner. Thus nominal communication is minimax-necessary over the weakly communicating class: without it, a nominal trajectory can contain no information about a state that an adversary can make recurrent.

Within the positive communicating regime, Theorem 14 constructs a two-state family satisfying $\delta < \delta_f^*(P)$ with minimax regret $\widetilde{\Theta}(T^{2/3})$. Thus Theorem 14 rules out a uniform square-root theorem even though the structural threshold is satisfied with a strict margin. □

Proposition 20. *Suppose $f \in \mathcal{F} \setminus (\mathcal{F}_{\mathrm{nc}} \cup \mathcal{F}_{\mathrm{ne}})$ and $\mathcal{G}(P) \in \mathcal{M}$. Then $\delta_f^{\mathrm{bell}}(P) = +\infty$. If*

$$C(P) = \mathcal{S},$$

then Theorem 9 gives $\mathrm{Reg}_{\mathrm{pol}}(T) = o(T)$ at every finite radius, without $c_{\mathrm{cov}}(P) < +\infty$. If $0 < \delta < \theta_f(1)$ and nominal communication is dropped, Proposition 14 gives a linear hidden-recurrence lower bound, even for examples satisfying $\delta < \delta_f^*(P)$. At $\delta \geq \theta_f(1)$, Theorem 5(iii) instead gives zero regret for every nominal kernel. If $\delta \geq \delta_f^\dagger(P)$, Proposition 15 gives the exact obstruction to memberwise weak communication and memberwise scalar nominal gains, while the robust Bellman equation remains scalar. Within the nonsaturated regime, Proposition 13 supplies the unavoidable finite-sample cost of deciding which transitions remain undeletable, and Theorem 14 supplies a $\widetilde{\Theta}(T^{2/3})$ creation subproblem with a strict common-core margin. Propositions 11 and 12 supply the same matched exponent separately in the two annuli above $\delta_f^*(P)$. Thus the horizon exponent is $2/3$ throughout the nonsaturated communicating creation regime $0 < \delta < \theta_f(1)$; at and above $\theta_f(1)$, Theorem 5(iii) gives zero regret.

Proof. Because $f \notin \mathcal{F}_{\mathrm{ne}}$, Theorem 3 gives $\delta_f^{\mathrm{bell}}(P) = +\infty$. If $C(P) = \mathcal{S}$, every state–action row is repeatedly sampled by the uniform exploration policy. These are precisely the assumptions of Theorem 9 for $f \notin \mathcal{F}_{\mathrm{ne}}$, so Theorem 9 gives $o(T)$ regret. It does not require $\delta < \delta_f^*(P)$ or finite coverage.

If $\delta \geq \theta_f(1)$, the sharper conclusion follows directly from Theorem 5(iii): every row ambiguity set is the full simplex and a reward-greedy selector has zero regret, regardless of $C(P)$. Hence all remaining lower-bound statements concern $0 < \delta < \theta_f(1)$.

Since $f \notin \mathcal{F}_{\mathrm{nc}}$, Proposition 13 applies. Since $f \notin \mathcal{F}_{\mathrm{ne}}$, creation is possible. The hard instances in Proposition 14 additionally require $\delta < \theta_f(1)$; their constructed robust cores are weakly communicating, so they satisfy $\delta < \delta_f^*(P_i)$. They have $C(P_i) = \{s\} \neq \mathcal{S}$ and linear regret despite satisfying $\delta < \delta_f^*(P_i) \leq \delta_f^\dagger(P_i)$. Hence the communication assumption cannot be removed uniformly. Independently, Proposition 15 proves the exact memberwise structural limit at $\delta_f^\dagger(P)$, while Proposition 13 quantifies the finite-sample testing cost that remains below that threshold. Finally, Theorem 14 applies to this class whenever $0 < \delta < \theta_f(1)$ and gives a communicating two-state $\widetilde{\Theta}(T^{2/3})$ subproblem satisfying $\delta < \delta_f^*(P)$. This lower bound isolates edge creation rather than deletion as the source of the larger exponent. Proposition 12 uses two individually deletable exits with undeletable aggregate mass to locate the same experiment in $\delta_f^*(P) \leq \delta < \delta_f^\dagger(P)$. Proposition 11 instead adds a closable action and locates it in $\delta_f^\dagger(P) \leq \delta < \delta_f^{\mathrm{bell}}(P) = +\infty$. Both keep all family constants bounded. □