SPAN-BASED ANALYSIS OF DISTRIBUTIONALLY ROBUST AVERAGE-REWARD MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study finite distributionally robust Markov games under the average-reward criterion, where each player evaluates the worst-case average-reward of policies against a player-specific, state–joint-action rectangular ambiguity set. We introduce the uniform optimal-response bias span H_{R} and develop studies on both equilibrium existence and learning guarantees in terms of this single parameter. Firstly, we show $H_{\mathrm{R}} < \infty$ under different chain structures and uncertainty sets, and that a finite H_{R} guarantees the existence of a stationary statewise robust average-reward Nash equilibrium. We then establish the convergence of Nash equilibria under discounted reward to average reward as the discount factor approaching 1, and develop a reduction framework to connect them. Furthermore, under total-variation ambiguity and balanced nominal generative-model sampling, we show that a model-based discounted learner with any exact discounted Nash oracle uses $\widetilde{O}(SA_{\times}d_{\max}(H_{\mathrm{R}} + \sigma H_{\mathrm{R}}^2)/\varepsilon^2)$ samples to learn an ε-robust average-reward Nash equilibrium, where $S, A_{\times}$ are the state and joint action space sizes, and $d_{\max}$ is the largest next-state support size. This guarantee holds simultaneously for every exact empirical Nash oracle. A matching lower bound on two-player, two-action games further proves the support factor unavoidable for this balanced, every-oracle interface, implying the tightness of our upper bound.

1 INTRODUCTION

Multi-agent systems that operate continuously must coordinate decisions while maintaining performance over long time horizons. Average-reward Markov games (Filar & Vrieze, 2012; Guo & Yang, 2008; Hernández-Lerma & Lasserre, 2000; Guo & Hernández-Lerma, 2003; Zheng & Guo, 2024; Nowak, 1999; Iwase et al., 1976; Küenle & Schurath, 2003; Tanaka & Wakuta, 1977; Jaśkiewicz & Nowak, 2001) capture this objective by evaluating each player's long-run reward rate and seeking a policy profile from which no player benefits by deviating. However, in practice, due to the Sim-to-Real gaps induced by, e.g., noise or adversarial attacks, the transition model used to compute such a profile may differ from the deployment environment, changing both the players' payoffs and their incentives to deviate. Distributionally robust Markov games (DR-MGs) address this mismatch by evaluating each player against the worst-case transition model in a prescribed ambiguity set (Kardeş et al., 2011; Zhang et al., 2020a; Mcmahan et al., 2024). The resulting learning problem is to find an equilibrium for these worst-case long-run objectives from observations of the nominal dynamics.

However, most existing DR-MG theory provides equilibrium and learning guarantees for discounted models (Kardeş et al., 2011; Zhang et al., 2020a) and finite-horizon models (Ma et al., 2023; Shi et al., 2024b; Farhat et al., 2026). Average reward introduces an additional dependence on the limiting recurrent behavior induced jointly by the players and adverse transitions, making it much more challenging. For example, stationary Nash equilibria need not exist without structural assumptions (Filar & Vrieze, 2012; Roch & Wang, 2026). Recent work establishes their existence under conditions like irreducibility, or under fixed-opponent weak communication together with uniqueness of the optimal bias modulo constants (Roch & Wang, 2026), which yet can be restrictive. A central question thus arises: *how to express the required structure for equilibrium existence in a generic way?*

We formulate this requirement through the term of *uniform optimal-response bias span* H_{R}. Bias spans are established complexity measures in nominal average-reward control (Zurek & Chen, 2024b; Wang & Si, 2025; Yang et al., 2026). Here we take the smallest bias span among the constant-

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.

gain Bellman solutions of each response problem, then the supremum over players and stationary opponents. This uniformity will address a statistical dependence specific to DR-MGs: an empirical Nash oracle chooses the opponents from the same samples used to estimate their response problems. Based on it, we then develop a reduction framework to approximate average-reward robust Nash equilibria through discounted ones, and prove their existence. We further analyze the statistical complexity of model-based reduction approach under balanced nominal generative model sampling, deriving a tight complexity bound uniformly for any exact empirical discounted Nash equilibrium oracle. Our contributions are as follows.

Equilibrium existence without bias uniqueness and a discounted reduction. Under the condition $H_{\mathrm{R}} < \infty$, we prove that every cluster point of stationary discounted robust Nash equilibria as $\gamma \uparrow 1$ is a stationary statewise robust average-reward Nash equilibrium, valid against history-dependent unilateral deviations (Theorem 3.3). The argument handles opponents that vary along the equilibrium sequence and does not rely on a unique bias, removing the bias-uniqueness requirement from the weak-communication regime of prior work (Roch & Wang, 2026). We further develop a reduction framework showing that the Nash gap under average-reward can be bounded by the one under discounted reward: $\mathrm{Gap}_{\mathrm{av}}(\pi) \le (1-\gamma)[H_{\mathrm{R}} + \mathrm{Gap}_{\gamma}(\pi)]$ (Theorem 3.4), so a sufficiently accurate discounted equilibrium with large discount factor gives an average-reward equilibrium guarantee.

Structural conditions that certify bounded span. We further explore different structural conditions under which the span is bounded uniformly over stationary opponents. Our certificates cover chain communication, hitting times, minorization, and recurrent-class overlap, including some multichain models, explicitly implying the applicability and feasibility of our studies. We also study the span boundedness under distributional divergence balls, where we show total variation and reverse KL-divergence balls with positive radii ensure a bounded span, and develop further studies under other divergence balls (Section 4 and Appendix F.4). We also give a reward-uniform characterization through disjoint closed regions, including quantitative residual-margin bounds and a reduction to deterministic opponents (Appendices F.4 and F.3).

Sample complexity uniform over empirical Nash oracles. With known rewards and total variation (TV) uncertainty sets whose radii are at most $\sigma \le 1/2$ and for some known $H \ge H_{\mathrm{R}}$, we prove a balanced model-based learner using any exact discounted Nash oracle achieves an ε-robust average-reward Nash equilibrium with high probability using

$$N_{\mathrm{tot}} = \widetilde{O}\left(S A_{\times} d_{\max} \frac{H + \sigma H^2}{\varepsilon^2} \right)$$

nominal transition samples (Theorem 5.1). Here S is the number of states, $A_{\times}$ is the number of joint actions, and $d_{\max} \le S$ is the largest *nominal* successor support. This guarantee holds simultaneously for every exact empirical oracle. We obtain this simultaneous guarantee by following robust response values along a segment of nominal centers and controlling both the optimal-response path and the selected-policy evaluation path using the population optimal-response span.

Tightness of the oracle interface and the role of selection. An adaptive control embedding gives the lower bound $\Omega(SA_{\times}(H + \sigma H^2)/\varepsilon^2)$ on action-budget classes that permit one active player (Theorem 6.1). We further construct a hard instance with two players with two actions each, showing a lower bound of $\Omega(Sd_{\max}(H + \sigma H^2)/\varepsilon^2)$ for balanced guarantees required to ensure every exact empirical discounted Nash oracle gives an ε-gap (Theorem 6.2). Thus the additional support factor is necessary for the stated oracle interface, including the dense rate $\widetilde{\Theta}(S^2(H + \sigma H^2)/\varepsilon^2)$ when $d_{\max} = \Theta(S)$. The quantifier over oracles is essential: Proposition I.3 and Proposition I.4 constructs different exact empirical oracle with state-linear sample complexity on the same hard family and achieve a state-linear span-dependent rate for shared-reward, shared-ambiguity team games under its unichain condition. The results therefore identify the hardness of learning in average-reward DR-MGs, and the cost of ensuring accuracy over arbitrary discounted equilibria.

2 Problem Formulation and Preliminaries

Finite games and stationary policies. We consider a finite distributionally robust Markov game (DR-MG) $\mathcal{M}$ with players $\mathcal{N} = [n]$, state space $\mathcal{S}$ of cardinality $S = |\mathcal{S}|$, and action space $\mathcal{A}_i$ of cardinality $A_i = |\mathcal{A}_i|$ for player i. Let $\mathcal{A} := \prod_{i \in \mathcal{N}} \mathcal{A}_i$ be the joint action space, with cardinality $A_{\times} := |\mathcal{A}| = \prod_{i \in \mathcal{N}} A_i$, and let $\mathcal{A}_{-i} := \prod_{j \ne i} \mathcal{A}_j$ be the joint action space of all players except i.

The nominal transition kernel is $P^0 = \{P^0_{s,\boldsymbol{a}} \in \Delta(\mathcal{S}) : (s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}\}$. At time t, the players observe state S_t and choose their actions simultaneously, forming $\boldsymbol{A}_t = (A_{1,t}, \ldots, A_{n,t})$. Player i receives the known reward $r_i(S_t, \boldsymbol{A}_t)$, where $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$; the next state is drawn from the transition row associated with $(S_t, \boldsymbol{A}_t)$. A stationary policy of player i is a map $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$. Define $\Pi_i := \prod_{s\in\mathcal{S}} \Delta(\mathcal{A}_i)$, $\Pi := \prod_{i\in\mathcal{N}} \Pi_i$, and $\Pi_{-i} := \prod_{j\neq i} \Pi_j$. For a stationary profile $\pi \in \Pi$, the players randomize independently conditional on the current state, so $\pi(\boldsymbol{a} \mid s) = \prod_i \pi_i(a_i \mid s)$. We write π_{-i} for the opponents' profile and use $\theta \in \Pi_{-i}$ when fixing the opponents of player i; in particular, $\theta(\boldsymbol{a}_{-i} \mid s) := \prod_{j\neq i} \theta_j(a_j \mid s)$. The notation (μ_i, θ) denotes the profile obtained by combining player i's response $\mu_i \in \Pi_i$ with these opponents. Under a history-dependent unilateral deviation, the opponents still draw fresh independent actions according to θ at each time.

$(s, \boldsymbol{a})$-Rectangular transition uncertainty. Player i evaluates a profile against its own ambiguity set, which captures the potential model mismatches and is centered at the nominal kernel P^0:

$$\mathcal{P}_i(P^0) := \prod_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{i,s,\boldsymbol{a}}(P^0_{s,\boldsymbol{a}}), \tag{1}$$

where each local set is a nonempty compact subset of $\Delta(\mathcal{S})$. This *state–joint-action rectangularity* means that an admissible kernel can be assembled by choosing its rows independently across $(s, \boldsymbol{a})$, and this model is the most standard model for DR-MGs (Shi et al., 2024b; Kardeş et al., 2011; Farhat et al., 2026; Roch & Wang, 2026). We call each such pair a *primitive row*. Nature observes the realized joint action before applying its corresponding transition row. Thus nature's minimization is performed separately for each joint action, before averaging over action randomization. Note that the worst-case evaluation is also performed separately for each player: the kernels minimizing different players' payoffs may differ, even if they use the identical ambiguity set $\mathcal{P}_i = \mathcal{P}_j$ but different local rewards (Shi et al., 2024b; Roch & Wang, 2026; Ma et al., 2023).

Our statistical results specialize to total-variation (TV) ambiguity: for any nominal row $q \in \Delta(\mathcal{S})$,

$$\mathcal{P}_{i,s,\boldsymbol{a}}(q) := \{p \in \Delta(\mathcal{S}) : \mathrm{TV}(p, q) \le \sigma_{i,s,\boldsymbol{a}}\}, \qquad \mathrm{TV}(p, q) := \tfrac{1}{2} \|p - q\|_1, \tag{2}$$

where the known radii satisfy $0 \le \sigma_{i,s,\boldsymbol{a}} \le \sigma \le 1/2$. Here σ is a common upper bound; the local radii may vary across players, states, and joint actions. These balls use the full simplex $\Delta(\mathcal{S})$: an admissible row may assign positive probability to a state outside the support of its center. The same simplex and radii are retained when the center is estimated from samples. The existence, span, and deterministic reduction results apply to the general nonempty compact rectangular sets in (1). When the center is omitted, write $\mathcal{P}_i := \mathcal{P}_i(P^0)$ and $\mathcal{P}_{i,s,\boldsymbol{a}} := \mathcal{P}_{i,s,\boldsymbol{a}}(P^0_{s,\boldsymbol{a}})$.

Robust average reward and the equilibrium objective. For a stationary profile π and stationary kernel P, let $\mathbb{E}^{\pi,P}_s$ denote expectation for the trajectory initialized at $S_0 = s$, with actions drawn according to π and transitions according to P. The robust average payoff of player i is

$$\underline{g}^{\pi}_i(s) := \inf_{P\in\mathcal{P}_i} \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi,P}_s \left[\sum_{t=0}^{T-1} r_i(S_t, \boldsymbol{A}_t)\right]. \tag{3}$$

The infimum ranges over stationary kernels. For each fixed (π, P), the expected Cesaro average has a limit because the induced chain is finite; the payoff may still depend on the initial state. For fixed opponents θ, define the optimal stationary-response gain and the statewise robust Nash gap by

$$\rho^\star_i(\theta, s) := \sup_{\mu_i\in\Pi_i} \underline{g}^{(\mu_i,\theta)}_i(s), \quad \mathrm{Gap}_{\mathrm{av}}(\pi) := \max_{i\in\mathcal{N}} \max_{s\in\mathcal{S}} \left[\rho^\star_i(\pi_{-i}, s) - \underline{g}^{\pi}_i(s)\right]. \tag{4}$$

A stationary profile π with $\mathrm{Gap}_{\mathrm{av}}(\pi) \le \varepsilon$ is called a statewise ε-robust average-reward Nash equilibrium. The case $\varepsilon = 0$ gives an exact equilibrium. Equivalently, every player can improve its own worst-case average reward by at most ε through any stationary unilateral deviation, from every initial state. Appendix C.2 also defines the payoff against history-dependent nature, whose row at each time may depend on the past and the realized joint action while remaining in the ambiguity set.

Discounted response problems and Bellman operators. For $\gamma \in (0, 1)$, define the *unnormalized* robust discounted value and its optimal response by

$$V^{\pi}_{i,\gamma}(s) := \inf_{P\in\mathcal{P}_i} \mathbb{E}^{\pi,P}_s \left[\sum_{t=0}^{\infty} \gamma^t r_i(S_t, \boldsymbol{A}_t)\right], \quad V^{\star,\theta}_{i,\gamma}(s) := \sup_{\mu_i\in\Pi_i} V^{\mu_i,\theta}_{i,\gamma}(s), \quad V^{\mu_i,\theta}_{i,\gamma} := V^{(\mu_i,\theta)}_{i,\gamma}.$$

These values lie in $[0, (1-\gamma)^{-1}]$. For fixed opponents, rectangularity gives the response operator

$$(T^{\theta}_{i,\gamma} v)(s) := \max_{a_i \in \mathcal{A}_i} \sum_{\boldsymbol{a}_{-i} \in \mathcal{A}_{-i}} \theta(\boldsymbol{a}_{-i} \mid s) \big[r_i(s, a_i, \boldsymbol{a}_{-i}) + \gamma \min_{p \in \mathcal{P}_{i,s,(a_i,\boldsymbol{a}_{-i})}} p^{\top} v \big]. \tag{5}$$

The maximum can be taken over pure actions because the expression is linear in player i's current action distribution. Replacing this maximum by averaging under $\mu_i(\cdot \mid s)$ gives the fixed-response operator $T^{\mu_i,\theta}_{i,\gamma}$. Both operators are monotone γ-contractions in $\|\cdot\|_\infty$, with unique fixed points $V^{\star,\theta}_{i,\gamma}$ and $V^{\mu_i,\theta}_{i,\gamma}$, respectively; this is the standard robust discounted Bellman characterization (Kardeş et al., 2011). The discounted robust Nash gap is

$$\mathrm{Gap}_\gamma(\pi) := \max_{i \in \mathcal{N}} \left\| V^{\star,\pi_{-i}}_{i,\gamma} - V^{\pi}_{i,\gamma} \right\|_\infty . \tag{6}$$

The difference is coordinatewise nonnegative. A stationary discounted robust Nash equilibrium has $\mathrm{Gap}_\gamma(\pi) = 0$, and an e_γ-approximate equilibrium has $\mathrm{Gap}_\gamma(\pi) \le e_\gamma$, where $e_\gamma \ge 0$.

Constant-gain Bellman equation. Let T^{θ}_i denote the operator in (5) with γ setting to one, the constant-gain Bellman equation for the robust average-reward is

$$\rho \mathbf{1} + h = T^{\theta}_i h, \qquad (\rho, h) \in \mathbb{R} \times \mathbb{R}^S, \tag{7}$$

where ρ is a scalar gain and h is a bias vector. Adding a constant to h preserves the equation, so its span measures the variation relevant to our analysis. The next section defines the optimal-response span parameter through these solutions. Under some structure assumptions (Roch et al., 2025; Wang & Si, 2025), the optimal gain $\rho^{\star}_i(\theta, s)$ is a constant over s and solves the Bellman equation. We further prove that this solution is also optimal against history-dependent admissible nature; the precise verification statements appear in Appendix C.2.

Sampling and the learning interface. In the learning setting, the learner knows the rewards and ambiguity radii and accesses the unknown nominal kernel through a generative model: querying a primitive row $(s, \boldsymbol{a})$ returns an independent next state drawn from $P^0_{s,\boldsymbol{a}}$. Our model-based learner uses a balanced experiment with N independent samples from each row, also independent across rows. The empirical nominal kernel $\widehat{P}^0$ consists of the corresponding next-state frequencies, and the total transition-sample count is $N_{\mathrm{tot}} = SA_{\times}N$. For accuracy $\varepsilon > 0$ and failure probability $\delta \in (0,1)$, the goal is to return a stationary profile $\widehat{\pi}$ satisfying $\mathbb{P}(\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \le \varepsilon) \ge 1 - \delta$, where the gap is evaluated in the population game centered at P^0 and the probability is over the samples and any learner randomization. The learner calls an exact discounted Nash oracle for the empirical robust game centered at $\widehat{P}^0$, retaining the original radii. An oracle is any measurable rule selecting an exact stationary equilibrium of that game. The upper bound holds simultaneously for every such oracle. Oracle computation is excluded from the transition-sample budget.

3 Optimal-Response Span: Existence and Discounted Reduction

In this section, we introduce the key term in our analysis, the optimal-response span, under which we develop a reduction-based framework for average-reward DR-MGs.

Definition 3.1 (Optimal-response span)**.** *Fix player i and stationary opponents θ. Let $\mathfrak{B}_i(\theta)$ be the set of constant-gain Bellman solutions (ρ, h) for the induced robust response problem. With the convention* $\inf \varnothing = +\infty$ *(i.e., when the constant-gain equation has no solution), define*

$$H_i(\theta) := \inf_{(\rho,h) \in \mathfrak{B}_i(\theta)} \max\{1, \mathrm{sp}(h)\}, \qquad H_{\mathrm{R}} := \max_i \sup_{\theta \in \Pi_{-i}} H_i(\theta). \tag{8}$$

Interpretation of H_{R}. The bias span measures the largest state-to-state transient advantage before the common long-run gain dominates, which has been identified as the key term in robust average-reward learning (Roch et al., 2025; 2026; Yang et al., 2026). Taking the minimum over all Bellman solutions avoids requiring unique bias, while the supremum over θ makes the parameter stable under a data-dependent equilibrium output. We highlight that this supremum is essential, and a span bound imposed only at population equilibria does not supply this uniform response control: empirical opponents need not be population-equilibrium opponents. Appendix C gives an explicit example with bounded equilibrium span and arbitrarily large off-equilibrium response span.

Our study of average-reward DR-MGs is based on the bounded span assumption.

Assumption 3.2. *The uniform optimal-response span satisfies $H_{\mathrm{R}} < \infty$. For the learning guarantee, a finite upper bound $H \geq H_{\mathrm{R}}$ is supplied to the learner.*

Knowledge of an optimal-span bound is also used in reduction-based single-agent learning (Roch et al., 2025; Yang et al., 2026). Note the definition implies $\mathfrak{B}_i(\theta) \neq \varnothing \iff H_i(\theta) < \infty$, so finite H_{R} requires constant-gain Bellman solvability for every response problem and a uniform bound on the minimum bias spans. It requires no uniqueness of the bias, response, or equilibrium. In particular, the weak-communication existence theorem of (Roch & Wang, 2026) imposes uniqueness of the optimal bias modulo constants, whereas our bounded span does not require this. The supremum over opponents supplies the response control needed at the data-dependent output of the empirical oracle. Also note finite H_{R} requires a state-independent robust optimal-response gain for every stationary opponent profile, and problems with state-dependent optimal gains require a different framework.

With the bounded-span assumption, we then develop the reduction framework, which connects the average-reward DR-MG with the discounted-reward one. We will first show that, under the assumption, we can construct an average-reward Nash equilibrium through discounted ones, therefore prove the existence of average-reward Nash equilibria. We then have the following results.

Theorem 3.3. *Under Assumption 3.2, every cluster point of stationary discounted robust Nash equilibria as $\gamma \uparrow 1$ is a stationary statewise robust average-reward Nash equilibrium, valid against arbitrary history-dependent unilateral deviations. Consequently, a robust average-reward Nash equilibrium exists under Assumption 3.2. (Details in Appendix C and Theorem D.4.)*

Proof sketch. The argument has three steps, and we define the discounted span parameter as $D_{\mathrm{R}} := \max\left\{1, \max_i \sup_{\theta\in\Pi_{-i}} \sup_{\gamma\in(0,1)} \mathrm{sp}\left(V_{i,\gamma}^{\star,\theta}\right)\right\}$. First, normalize a minimum-span constant-gain solution h by $\min_s h(s) = 0$. It yields discounted sub- and supersolutions $\rho(1-\gamma)^{-1}\mathbf{1} + h - \mathrm{sp}(h)\,\mathbf{1}$ and $\rho(1-\gamma)^{-1}\mathbf{1} + h$. Monotonicity and contraction sandwich the discounted optimum and give $D_{\mathrm{R}} \leq 2H_{\mathrm{R}}$. Second, if discounted spans are uniformly bounded, subtracting the minimum coordinate makes the values precompact. Any vanishing-discount limit solves the constant-gain equation, proving $H_{\mathrm{R}} \leq D_{\mathrm{R}}$. Third, choose a discounted equilibrium sequence. Compactness gives a convergent policy subsequence and playerwise normalized value limits. Passing the normalized discounted Bellman equations and greedy equalities to the limit yields one gain–bias witness per player. Rowwise lower and upper Bellman inequalities then telescope against arbitrary unilateral histories. □

The same fixed-opponent comparison, together with a one-sided fixed-policy Bellman bound, yields the following Nash-gap reduction. Its uniform formulation allows the profile to be chosen by an empirical equilibrium oracle. The detailed proof is developed in Theorem E.3 in appendix.

Theorem 3.4. *For every stationary profile π and any $\gamma \in (0,1)$, it holds that*

$$\mathrm{Gap}_{\mathrm{av}}(\pi) \leq (1-\gamma)\left[H_{\mathrm{R}} + \mathrm{Gap}_{\gamma}(\pi)\right]. \tag{9}$$

Therefore, for any $H \geq H_{\mathrm{R}}$ and $0 < \varepsilon \leq 1$, choosing $\gamma = 1 - \varepsilon/(2H)$ and any profile with $\mathrm{Gap}_{\gamma}(\pi) \leq H$ implies $\mathrm{Gap}_{\mathrm{av}}(\pi) \leq \varepsilon$.

Consequently, choosing $1-\gamma = \varepsilon/(2H)$ makes an unnormalized discounted Nash gap of at most H sufficient for an ε-average-reward guarantee. The learning analysis below controls the population discounted gap of an empirical oracle output at this accuracy.

4 When Is the Optimal-Response Span Bounded?

We give two complementary ways to certify $H_{\mathrm{R}} < \infty$: communication properties of the fixed-opponent response models, and conditions on their divergence balls. Both establish scalar-gain Bellman solvability and bias bounds uniform over stationary opponents. Write $R_i := \max_{s,\boldsymbol{a}} r_i(s,\boldsymbol{a}) - \min_{s,\boldsymbol{a}} r_i(s,\boldsymbol{a})$ and $R_{\mathrm{osc}} := \max_i R_i$.

Conditions on response chains. Our first direction is through the response chain: since bounded span is related to the constant-gain Bellman equation solvability, it can be ensured by the structures of the induced chains by opponent policies. Specifically, fix player i and stationary opponents θ. The sets of admissible transition rows for player i induced by θ are $\mathcal{Q}_{s,a_i}^{i,\theta} \triangleq \left\{ \sum_{a_{-i}} \theta(a_{-i} \mid s) p_{a_{-i}} : p_{a_{-i}} \in \right.$

$\mathcal{P}_{i,s,(a_i,a_{-i})}\}$. A *full row selection* $q = (q_{s,a_i})$ chooses one admissible row for every state and own action. Together with agent i's stationary response $\mu : \mathcal{S} \to \Delta(\mathcal{A}_i)$, it induces a Markov chain with transition matrix $Q^{\mu,q}(s,\cdot) = \sum_{a_i} \mu(a_i \mid s) q_{s,a_i}$. Following (Puterman, 2014), a chain is *unichain* if it has exactly one closed irreducible recurrent class with a possibly empty set of transient states. A controlled MDP is *weakly communicating* if it has a nonempty core whose states are mutually reachable under suitable stationary responses, while every state outside that core is transient under every stationary response.

We then derive the following sufficient conditions from chain structures.

Theorem 4.1. *Suppose that, for every player i and stationary opponents θ, at least one of the following holds. Then, for every bounded reward specification and every i, θ, the equation $\rho_i(\theta)\mathbf{1} + h_i^\theta = T_i^\theta h_i^\theta$ admits a solution, and the minimum bias spans are uniformly bounded over i, θ, hence $H_{\mathrm{R}} < \infty$:*

1. ***All selections are unichain.*** *For every deterministic stationary response a and every full row selection q, $Q^{a,q}$ is unichain.*
2. ***Nature can supply an irreducible chain.*** *For each deterministic stationary response a, there exists a full row selection $q = q(a)$ such that $Q^{a,q}$ is irreducible.*
3. ***The player can supply an irreducible chain.*** *For each full row selection q, there exists a stationary response $\mu = \mu(q)$ such that $Q^{\mu,q}$ is irreducible.*
4. ***Every selected controlled model is weakly communicating.*** *For each full row selection q, the MDP with action rows q_{s,a_i} is weakly communicating. Its core and communicating responses may depend on i, θ, q.*

The proofs are deferred to Appendix F.1. These certificates specialize one-sided communication criteria from robust control (Wang & Si, 2025, Corollaries 7.1–7.3) and make their span control uniform over opponents. Condition 1 permits transient states, unlike the all-pair irreducibility regime of (Roch & Wang, 2026). Conditions 2 and 3 allow the irreducible witness to depend on the fixed response or row selection, respectively. Condition 4 permits the communicating core to vary across selected models. The resulting existence guarantee also removes the bias-uniqueness requirement in the weak-communication regime of (Roch & Wang, 2026).

Divergence ambiguity. We also investigate the span boundedness for the widely used distributional divergence balls in distributionally robust RL, e.g., (Shi et al., 2023; Panaganti & Kalathil, 2022; Roch et al., 2025; Yang et al., 2026), specified as $\mathcal{P}_{i,s,\boldsymbol{a}}(q) := \{p \in \Delta(\mathcal{S}) : D(p,q) \le \sigma_{i,s,\boldsymbol{a}}\}$, for some divergence D like total variation or KL-divergence, and some nominal center kernel q. In the following, we first show that positive-radius full-simplex TV and reverse-KL balls ensure bounded spans. Studies for other divergence balls are also developed and deferred to Appendix F.4.

Theorem 4.2. *Write $q = P^0_{s,\boldsymbol{a}}$ and $\underline{\kappa}_i := \min_{s,\boldsymbol{a}} \kappa_{i,s,\boldsymbol{a}}$. In each case below, the stated row conditions hold for every $i, s, \boldsymbol{a}$. Each case guarantees, for every player i and stationary opponents θ, a scalar-gain Bellman solution $\rho_i(\theta)\mathbf{1} + h_i^\theta = T_i^\theta h_i^\theta$ and $H_{\mathrm{R}} < \infty$.*

*(A) **Total variation.** Suppose $\mathcal{P}_{i,s,\boldsymbol{a}} = \{p \in \Delta(\mathcal{S}) : \mathrm{TV}(p,q) \le \sigma_{i,s,\boldsymbol{a}}\}$, where $0 \le \sigma_{i,s,\boldsymbol{a}} \le 1$ and $\underline{\sigma}_i := \min_{s,\boldsymbol{a}} \sigma_{i,s,\boldsymbol{a}} > 0$ for every i. Then $\mathrm{sp}(h_i^\theta) \le \frac{R_i}{\underline{\sigma}_i}$, and $H_{\mathrm{R}} \le \max\{1, \max_i R_i/\underline{\sigma}_i\}$.*

*(B) **Reverse KL.** Suppose $\mathcal{P}_{i,s,\boldsymbol{a}} = \{p \in \Delta(\mathcal{S}) : D_{\mathrm{KL}}(q\|p) \le \kappa_{i,s,\boldsymbol{a}}\}$ and $\underline{\kappa}_i > 0$ for every i. Then $\mathrm{sp}(h_i^\theta) \le \frac{R_i}{1-e^{-\underline{\kappa}_i}}$, and $H_{\mathrm{R}} \le \max\{1, \max_i R_i/(1-e^{-\underline{\kappa}_i})\}$.*

Both cases let nature assign a uniformly positive probability to any target state, yielding bounds uniform over stationary opponents. Previous studies under these models additionally assume unichain or irreducible conditions (Roch et al., 2025; 2026; Yang et al., 2026), which we showed unnecessary here. All the proofs and details are presented in Appendix F.4.

Additional certificates. Appendices F.2 and F.3 develop hitting-time, reachability and contraction bounds, and a reward-uniform closed-region characterization following Akian et al. (2015). The certificates support equilibrium existence and the discounted-to-average reduction. The statistical analysis below specializes to TV balls.

5 LEARNING WITH A DISCOUNTED NASH ORACLE: TV BALLS

In this section, we study the sample complexity of solving an average-reward DR-MG through discounted approximation, specifically under the TV balls as (2). We focus on the most standard learning setting with a nominal non-adaptive sampling oracle: for any $(s, \boldsymbol{a})$ pair, the learner acquires N i.i.d. samples $s' \sim P^0_{s,\boldsymbol{a}}$ generated following nominal kernels (Agarwal et al., 2020; Panaganti & Kalathil, 2022; Shi et al., 2023). This setting reveals the most fundamental information-theoretic and statistical complexity in learning (Shi et al., 2024b; Zhang et al., 2020b; Li et al., 2022). Whether adaptive sampling can bypass this dependence is left for future work. We consider the plug-in model-based algorithm: for given H, ε, δ, set $\gamma = 1 - \varepsilon/(2H)$. Draw N independent samples from each $(s, \boldsymbol{a})$, form the empirical nominal kernel $\widehat{P}^0$, retain the original TV radii, and call an exact stationary γ-discounted robust Nash oracle to return its profile $\widehat{\pi}$[1] (e.g., from (Zhang et al., 2020a)).

Theorem 5.1. *Under TV balls (which implies Assumption 3.2), let $H \geq H_{\mathrm{R}}$, $0 < \varepsilon \leq 1$, and $0 < \delta < 1/2$. There is a universal constant C such that*

$$N \geq C d_{\max} \frac{H + \sigma H^2}{\varepsilon^2} \log\left(\frac{2S^2 A_\times}{\delta}\right), \text{ or } N_{\mathrm{tot}} = SA_\times N = \widetilde{O}\left(SA_\times d_{\max} \frac{H + \sigma H^2}{\varepsilon^2}\right), \quad (10)$$

implies $\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \leq \varepsilon$ with probability at least $1 - \delta$, simultaneously for every exact empirical Nash oracle. Here $d_{\max} := \max_{s,\boldsymbol{a}} |\operatorname{supp} P^0_{s,\boldsymbol{a}}| \leq S$ is the largest nominal successor support.

Note that such reduction complexity is also studied in robust/nominal average-reward MDPs (Roch et al., 2025; Yang et al., 2026; Zurek & Chen, 2024b), however, a further difficulty arises in games: even after the empirical kernel is fixed, different equilibria may induce different opponents and continuation values. Consequently, applying a single-agent guarantee separately to fixed opponents does not by itself control the response problems selected by an empirical Nash oracle.

Proof architecture. For a fixed stationary opponent profile, each player's deviation problem is a robust MDP. The profile relevant to the theorem, however, is $\widehat{\theta} = \widehat{\pi}_{-i}$, which is chosen by the empirical Nash oracle from the same transition samples used to construct the response problem. Conditioning on $\widehat{\theta}$ does not restore the product sampling law required by a fixed-MDP concentration theorem. The key proof device is therefore a deterministic perturbation theorem on one event that is uniform over the opponent profile and the equilibrium branch selected by the oracle.

Our proof has four main steps. First, coordinatewise Bernstein concentration yields one event that holds simultaneously for every primitive row and every continuation vector. If p is a nominal row, $\widehat{p}$ is its empirical estimate, and $d(p) = |\operatorname{supp}(p)|$, then, with $\ell = \log(2S^2A_\times/\delta)$, we have $\left|(\widehat{p} - p)^\top v\right| \leq \sqrt{\frac{2d(p)\ell}{N} \operatorname{Var}_p(v)} + \frac{2d(p)\ell}{3N} \operatorname{sp}(v)$ for every $v \in \mathbb{R}^S$. The quantifier over all v makes the event valid after the oracle has selected both the opponents and the equilibrium branch. TV duality expresses each robust expectation as a maximization over scalar clipping thresholds. Since clipping cannot increase span or variance, the same event controls all robust Bellman queries and all data-dependent clipping thresholds. Jensen's inequality and the law of total variance handle opponent and response mixtures without introducing an action-space factor.

Second, after fixing an arbitrary opponent profile on this simultaneous event, we interpolate every nominal row along $p_x(t) = (1-t)p_x + t\widehat{p}_x$, $0 \leq t \leq 1$. The corresponding optimal-response and fixed-response values are globally Lipschitz and piecewise rational in t. On an interval with fixed maximizing actions and fixed active bases of the TV linear programs, differentiating the Bellman equality gives $v_t' = \gamma(I - \gamma Q_t)^{-1} b_t$, where Q_t is the current Bellman-minimizing transition matrix and b_t contains the row-center perturbations. This center-path construction follows response values while avoiding any differentiation or regularity assumption on the set-valued Nash correspondence, which may switch discontinuously among equilibrium branches.

Third, the resolvent is controlled by the Bellman equation satisfied by the *current* value and the *same* minimizing matrix on the path. If $B_t = \operatorname{sp}(v_t)$ and $\lambda = 1 - \gamma$, a squared-bias identity

[1] The same sample bound also suffices for an approximate oracle whose output is a stationary product profile with empirical unnormalized discounted Nash gap at most $H/4$. Explicitly, it is enough that $\max_i \max_s \left[\widehat{V}^{\star,\widehat{\pi}_{-i}}_{i,\gamma}(s) - \widehat{V}^{\widehat{\pi}}_{i,\gamma}(s)\right] \leq H/4$, where $\widehat{\cdot}$ denote values in the empirical game. Thus exact equilibrium computation is unnecessary for the statistical guarantee.

yields $\gamma \left\| (I - \gamma Q_t)^{-1} \sqrt{\mathrm{Var}_{Q_t}(v_t)} \right\|_\infty \le \frac{\sqrt{2B_t}}{\lambda}$. Combining this identity with the simultaneous row event and the comparison between nominal and TV-adverse variances gives, for $a = d_{\max}\ell/N$, $\|v_t'\|_\infty \le \frac{2\sqrt{aB_t}+\sqrt{2a\sigma}\,B_t+5aB_t}{\lambda}$. With $B_t = O(H)$ and $\lambda = \varepsilon/(2H)$, the first term gives the H/ε^2 scale and the nominal-to-adverse variance comparison gives the additional $\sigma H^2/\varepsilon^2$ scale.

Finally, Assumption 3.2 bounds the population optimal response but does not directly bound intermediate empirical values or the value of the oracle-selected policy. A two-sided continuation argument closes this gap. Along the forward optimal-response path, the derivative bound rules out a first time at which the span reaches a larger localization radius and bounds the endpoint perturbation. At the empirical endpoint, exact Nash optimality identifies the oracle-selected response with an empirical optimal response. We then hold this response fixed and follow its policy-evaluation path backward to the population center; a second first-exit argument bounds its population value loss. The two paths give a population discounted response loss at most H, and Theorem 3.4 converts this into the claimed ε-average-reward Nash guarantee. Appendix G proves the path regularity, Bellman variance identity, and both continuation arguments. □

Why an arbitrary empirical equilibrium costs a support factor. The additional term $d_{\max}$ is not from the discounted-to-average reduction or our analysis; Instead, it arises because an empirical Nash oracle selects both the opponents and their continuation values from the same samples used to estimate the transition rows. A fixed-opponent robust-MDP bound cannot be applied merely by conditioning on those selected opponents, because this conditioning generally changes the row-sampling law.

For one nominal row p, put $\Delta := \widehat{p} - p$ and $d(p) := |\operatorname{supp} p|$. The exact identities $\sup_{\mathrm{Var}_p(v)\le\nu} |\Delta^\top v| = \left(\nu \sum_{x:p(x)>0} \frac{\Delta(x)^2}{p(x)}\right)^{1/2}$, and $\mathbb{E}\sum_{x:p(x)>0} \frac{\Delta(x)^2}{p(x)} = \frac{d(p)-1}{N}$ show why controlling all data-dependent continuation vectors has root-mean-square cost $\sqrt{\nu(d(p)-1)/N}$ (Proposition I.5). Our coordinatewise Bernstein event supplies the corresponding simultaneous upper bound. This identity alone is not a game lower bound: the maximizing vector need not be an equilibrium value. Theorem 6.2 supplies a realizable equilibrium construction and proves necessity for the balanced every-oracle guarantee.

Why existing decoupling arguments do not transfer automatically. Independent stage-specific samples separate continuation values from current-stage rows in finite-horizon games (Shi et al., 2024b). Single-agent discounted MDPs have a unique optimal value, which supports absorbing-state leave-one-out arguments (Agarwal et al., 2020; Yang et al., 2026). Ordinary zero-sum games likewise have a unique Shapley value; policy guarantees additionally require the planning properties in the corresponding result (Zhang et al., 2023). A general-sum empirical equilibrium can instead have several value branches, and an arbitrary oracle may choose its branch using the omitted row. A response problem being an MDP after opponents are fixed does not resolve this dependence. TV clipping and the deterministic reduction do not add the support factor in our analysis. More detailed discussions on the tightness are deferred to Appendix I. The next section further proves that this support factor is necessary for balanced guarantees covering every exact empirical oracle.

6 Lower Bounds and Tightness

The two proof families are separated in Appendix H: Appendix H.1 treats adaptive generative sampling, and Appendix H.2 treats balanced every-oracle guarantees. Their implications and the designed-oracle results are collected separately in Appendix I.

We first distinguish an adaptive control-type lower bound from the cost of a guarantee valid for every empirical Nash oracle under balanced sampling. Theorem H.2 proves the following bound through a two-successor control construction, a game embedding, and a change-of-measure argument that accounts for adaptive sample allocation.

Theorem 6.1 (Adaptive lower bound on an embedded control class)**.** *There are universal constants $c, c_0 > 0$ such that, for $S \ge 4$, $A_\times \ge 2$, $H \ge 4$, $0 \le \sigma \le 1/2$, and $0 < \varepsilon \le c_0$, a class with $H_{\mathrm{R}} \le H$ and nominal row supports at most two requires $\sup_{\mathcal{M}} \mathbb{E}_{\mathcal{M}}[T] \ge cSA_\times \frac{H+\sigma H^2}{\varepsilon^2}$ from every adaptive generative-model learner returning a stationary ε-robust average-reward Nash equilibrium*

with probability at least $3/4$ *on each instance. This class has one active player with* $A_\times$ *actions and remaining players with one action.*

The $A_\times$ conclusion concerns classes specified by a total joint-action budget whose allocation across players may vary. For a prescribed action-size vector, this control embedding gives $A_{\max} = \max_i A_i$ in place of $A_\times$. Thus, within the common parameter regime and up to logarithmic factors, the comparison with the available adaptive lower bound leaves a factor $d_{\max}$ on the action-budget class, and a factor $d_{\max}A_\times/A_{\max}$ for a general prescribed action-size vector. The latter comparison contains an unresolved action-space gap as well as the successor-support gap. Theorem 5.1 supplies the stronger requirement that, under balanced sampling, its guarantee holds for *every exact empirical discounted Nash oracle*. The following theorem isolates the cost of that requirement on a fixed-action class. We construct a two-player, two-action common-interest game and an exact empirical Nash oracle that aligns its continuation values with the sampling errors of a dense transition row. The selected profile satisfies every empirical Nash inequality, yet controlling its population Nash gap requires $N = \Omega(d_{\max}(H + \sigma H^2)/\varepsilon^2)$ samples per row. This establishes the successor-support cost of the balanced every-oracle interface. The joint construction and its all-budget completion are Theorem H.12 and Corollary H.14.

Theorem 6.2 (Joint lower bound for guarantees uniform over empirical Nash oracles)**.** *There are universal constants* $c, \varepsilon_0 > 0$ *such that the following holds for* $S \geq 4$*,* $3 \leq d_{\max} \leq S - 1$*,* $H \geq 1$*,* $0 \leq \sigma \leq 1/2$*, and* $0 < \varepsilon \leq \varepsilon_0$*. Set* $\gamma = 1 - \varepsilon/(2H)$*. Consider the class of two-player, two-action common-interest games whose maximum nominal row support is at most* $d_{\max}$*, whose robust response-span parameter satisfies* $H_{\mathrm{R}} \leq H$*, and whose full-simplex rectangular TV radii are at most* σ*. If a balanced empirical-model guarantee succeeds with probability at least* $3/4$ *on every game in this class for every exact measurable empirical discounted Nash oracle, then*

$$N \geq c\, d_{\max} \frac{H + \sigma H^2}{\varepsilon^2}, \quad N_{\mathrm{tot}} = 4SN = \Omega\left(S d_{\max} \frac{H + \sigma H^2}{\varepsilon^2} \right). \tag{11}$$

The necessity holds for every integer sample budget $N \geq 1$*.*

This lower bound matches our upper bound jointly in $S, d_{\max}, H, \sigma$, and ε, up to logarithmic factors, on the stated fixed-action class. In dense games it gives a quadratic state dependence for balanced guarantees covering every exact empirical oracle. Thus the upper bound is rate optimal, up to logarithmic factors, for this balanced every-oracle interface on the fixed-action class.

What can improve with discounted plug-in. The lower bound rules out removing $d_{\max}$ while retaining the every-oracle guarantee. It does not rule out an algorithm that specifies which equilibrium to select. On our hard family, a oracle that fixes all branch equilibria to the low continuation has $\mathbb{P}\{\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) > \varepsilon\} \leq e^{-16N\varepsilon^2}$, and is an exact empirical discounted equilibrium on every sample realization (Proposition I.3). Thus even this family admits a balanced discounted plug-in learner with $O(S\varepsilon^{-2}\log(1/\delta))$ samples. For the broader class with identical rewards and identical ambiguity sets, a centralized joint-action optimizer is an exact empirical Nash oracle. Under the additional unichain hypothesis of Proposition I.4, it achieves $\widetilde{O}(SA_\times(H + \sigma H^2)/\varepsilon^2)$ samples. These oracles achieve the improved rate on their respective subclasses. Whether a designed oracle attains the same rate under only the response-span assumption for general-sum games or player-specific ambiguity remains open. Appendix I.4 discusses the additional requirements.

7 Conclusion

In this work, we developed a span-based framework for distributionally robust average-reward Markov games. We showed that uniformly bounded optimal-response bias spans ensure stationary robust Nash equilibrium existence and yield a quantitative discounted-to-average reduction. Under total variation ambiguity, we proved the model-based learner can find an ε-Nash equilibrium with $\widetilde{O}(SA_\times d_{\max}(H + \sigma H^2)/\varepsilon^2)$ samples, uniformly over every empirical discounted Nash oracle. A matching lower bound for two players with two actions each establishes the successor-support cost of this uniform guarantee under balanced sampling. We also provided communication and divergence certificates for bounded span, with TV and reverse-KL results in the main paper and other-divergence extensions in the appendix. The remaining statistical question is whether a designed oracle achieves the state-linear benchmark under only the response-span assumption in general-sum games or with player-specific ambiguity, or if an unbalanced, adaptive learner can improve the complexity.

AI USE STATEMENT

In this work, we used generative AI tools to help review and revise proofs, identify potential gaps in arguments, and interpret the scope and implications of the results. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, generate synthetic datasets, implement research methods, translate content, or design experiments, and dataset cleaning and reformatting and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to assist with manuscript organization, language editing, and LaTeX preparation. We have reviewed all AI-assisted work. In particular, we checked the modified mathematical statements and proof arguments, verified cited sources and comparisons with prior work, and reviewed the revised text for accuracy and consistency. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work is theoretical and does not involve any external datasets or real-world deployment. The guarantees of DR-MG in this paper hold only under the stated modeling assumptions and within the specified ambiguity sets. Any practical deployment should separately consider domain-specific requirements concerning safety, fairness, privacy, security, and legal compliance.

REPRODUCIBILITY STATEMENT

This paper is theoretical and uses no external datasets or empirical benchmarks. The appendices provide complete proofs of the equilibrium-existence and discounted-reduction results, the sample-complexity upper bound, both lower-bound constructions, and the structural conditions for bounded span.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 67–83. PMLR, 2020.

Marianne Akian, Stéphane Gaubert, and Antoine Hochart. Ergodicity conditions for zero-sum games. *Discrete and Continuous Dynamical Systems*, 35(9):3901–3931, 2015. doi: 10.3934/dcds.2015.35.3901.

Peter Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 35–42. AUAI Press, 2009.

Jeremiah Birrell, Markos A Katsoulakis, and Yannis Pantazis. Optimizing variational representations of divergences and accelerating their statistical estimation. *IEEE Transactions on Information Theory*, 68(7):4553–4572, 2022.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. *Concentration Inequalities*, pp. 208–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_9. URL https://doi.org/10.1007/978-3-540-28650-9_9.

Krishnendu Chatterjee, Ehsan Kafshdar Goharshady, Mehrdad Karrabi, Petr Novotnỳ, and Đorđe Žikelić. Solving long-run average reward robust mdps via stochastic games. *arXiv preprint arXiv:2312.13912*, 2023.

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, 2025. URL https://openreview.net/forum?id=9Sb7YIHK4e.

Imre Csiszár and Paul C Shields. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Zain Ulabedeen Farhat, Debamita Ghosh, George K. Atia, and Yue Wang. Sample-efficient distributionally robust multi-agent reinforcement learning via online interaction. In *Proc. International Conference on Learning Representations (ICLR)*, 2026.

A. Federgruen. On n-person stochastic games by denumerable state space. *Advances in Applied Probability*, 10(2):452–471, 1978.

Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Arlington M Fink. Equilibrium in a stochastic n-person game. *Journal of science of the hiroshima university, series ai (mathematics)*, 28(1):89–93, 1964.

Stéphane Gaubert and Zheng Qu. Dobrushin ergodicity coefficient for markov operators on cones, and beyond. *arXiv preprint arXiv:1302.5226*, 2013.

Julien Grand-Clément and Marek Petrik. Reducing blackwell and average optimality to discounted mdps via the blackwell discount factor. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 52628–52647, 2023.

Julien Grand-Clément, Nian Si, and Shengbo Wang. Tractable robust markov decision processes. *arXiv preprint arXiv:2411.08435*, 2024.

Xianping Guo and Onésimo Hernández-Lerma. Zero-sum games for continuous-time markov chains with unbounded transition and average payoff rates. *Journal of Applied Probability*, 40(2):327–345, 2003.

Xianping Guo and Jie Yang. A new condition and approach for zero-sum stochastic games with average payoffs. *Stochastic analysis and applications*, 26(3):537–561, 2008.

Onésimo Hernández-Lerma and Jean B Lasserre. Zero-sum stochastic games in borel spaces: average payoff criteria. *SIAM Journal on Control and Optimization*, 39(5):1520–1539, 2000.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

Seiichi Iwase, Kensuke Tanaka, and Kazuyoshi Wakuta. On Markov games with the expected average reward criterion. *Science Reports of Niigata University. Series A, Mathematics*, 13:31–41, 1976. URL https://projecteuclid.org/journals/science-reports-of-niigata-university-series-a-mathematics/volume-13/issue-none/On-Markov-games-with-the-expected-average-reward-criterion/nihmj/1273779013.full.

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

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

Anna Jaśkiewicz and Andrzej S Nowak. On the optimality equation for zero-sum ergodic stochastic games. *Mathematical methods of operations research*, 54:291–301, 2001.

Yuchen Jiao and Gen Li. Minimax-optimal multi-agent robust reinforcement learning. *arXiv preprint arXiv:2412.19873*, 2024.

Erim Kardeş, Fernando Ordóñez, and Randolph W Hall. Discounted robust stochastic games and an application to queueing control. *Operations research*, 59(2):365–382, 2011.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Heinz-Uwe Küenle and Ronald Schurath. The optimality equation and ε-optimal strategies in markov games with average reward criterion. *Mathematical methods of operations research*, 56:451–471, 2003.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent rl in markov games with a generative model. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 15353–15367, 2022.

Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, July 2009. arXiv: 0907.3740.

Jeremy Mcmahan, Giovanni Artiglio, and Qiaomin Xie. Roping in uncertainty: Robustness and regularization in markov games. In *Proc. International Conference on Machine Learning (ICML)*, pp. 35267–35295. PMLR, 2024.

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

Andrzej S Nowak. Optimal strategies in a class of zero-sum ergodic stochastic games. *Mathematical methods of operations research*, 50:399–419, 1999.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. *arXiv preprint arXiv:2112.01506*, 2022.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1 edition, April 2014. ISBN 978-1-118-62587-3.

Zachary Roch and Yue Wang. Distributionally robust markov games with average reward. In *Proc. International Conference on Machine Learning (ICML)*, 2026.

Zachary Roch, George Atia, and Yue Wang. A reduction framework for distributionally robust reinforcement learning under average reward. In *Proc. International Conference on Machine Learning (ICML)*. PMLR, 2025.

Zachary Roch, George Atia, and Yue Wang. Model-free robust average-reward reinforcement learning with sample complexity analysis. In *Proc. International Conference on Machine Learning (ICML)*, 2026.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36:79903–79917, 2023.

Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Breaking the curse of multiagency in robust multi-agent reinforcement learning. *arXiv preprint arXiv:2409.20067*, 2024a.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Proc. International Conference on Machine Learning (ICML)*, pp. 44909–44959. PMLR, 2024b.

Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Policy optimization for robust average reward mdps. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 17348–17372, 2024.

Kensuke Tanaka and Kazuyoshi Wakuta. On continuous time Markov games with the expected average reward criterion. *Science Reports of Niigata University. Series A, Mathematics*, 14:15–24, 1977. URL `https://projecteuclid.org/journals/nihonkai-mathematical-journal/volume-14/issue-none/On-Continuous-Time-Markov-Games-With-The-Expected-Average-Reward/nihmj/1273779029.full`.

Qiuhao Wang, Yuqi Zha, Chin Pang Ho, and Marek Petrik. Provable policy gradient for robust average-reward mdps beyond rectangularity. In *Forty-second International Conference on Machine Learning*, 2025.

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

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward markov decision processes. In *International Conference on Learning Representations*, volume 2024, pp. 29843–29861, 2024.

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

Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 36431–36469. PMLR, 2023b.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2505–2513, 2010.

Yang Xu, Swetha Ganesh, and Vaneet Aggarwal. Efficient q-learning and actor-critic methods for robust average reward reinforcement learning. *arXiv preprint arXiv:2506.07040*, 2025a.

Yang Xu, Washim Uddin Mondal, and Vaneet Aggarwal. Finite-sample analysis of policy evaluation for robust average reward reinforcement learning. *arXiv preprint arXiv:2502.16816*, 2025b.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 9728–9754. PMLR, 2023.

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

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

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020a.

Kaiqing Zhang, Sham M Kakade, Tamer Basar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24(175):1–53, 2023.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *arXiv preprint arXiv:2006.03864*, 2020b.

Zewu Zheng and Xin Guo. Zero-sum non-stationary stochastic games with the long-run average criterion. *Applied Mathematics & Optimization*, 90(2):43, 2024.

Matthew Zurek and Yudong Chen. The plug-in approach for average-reward and discounted mdps: Optimal sample complexity analysis. *arXiv preprint arXiv:2410.07616*, 2024a.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward mdps. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

APPENDIX ORGANIZATION

The appendix follows the order of the main paper. The table below maps each main-paper section to its supporting definitions, results, and proofs.

Appendix	Purpose	Main-paper section
A	Related work	1
B	Model and learning interfaces	2
C	Optimal-response span and Bellman verification	3
D	Equilibrium existence	3
E	Deterministic discounted-to-average reduction	3
F	Conditions that certify bounded span	4
G	Statistical upper bound	5
H	Statistical lower bounds	6
I	Tightness, oracle design, and open questions	6

The two lower-bound arguments are treated separately: Appendix H.1 proves the adaptive generative-model lower bound, and Appendix H.2 proves the lower bound for balanced guarantees covering every exact empirical Nash oracle. Supporting lemmas and constructions appear within the corresponding subsection. The interpretation of these bounds and the oracle-design results are collected in Appendix I.

A RELATED WORK

Our results connect average-reward stochastic games, robust dynamic programming, and statistical learning with a generative model. We organize these connections by the mathematical problem being solved: equilibrium existence, structural conditions for long-run control, planning with model uncertainty, and learning from an estimated transition model. In comparing learning guarantees, we distinguish the uncertainty set, sampling access, equilibrium concept, and requirements on the empirical planning oracle.

A.1 CLASSICAL STOCHASTIC GAMES AND AVERAGE-REWARD EQUILIBRIA

Shapley (1953) established the dynamic programming foundation for discounted zero-sum stochastic games, including the value and stationary optimal strategies. Fink (1964) proved stationary equilibrium existence for finite discounted general-sum stochastic games. Discounting controls the contribution of the distant future; the average-reward problem instead requires control of the long-run transition structure. Classical treatments develop this connection through recurrence and communication conditions (Filar & Vrieze, 2012). For non-cooperative general-sum games, Federgruen (1978) established average-payoff equilibria under recurrence conditions and obtained them as limits of discounted equilibria. In the zero-sum setting, Nowak (1999) proved stationary optimality under geometric ergodicity conditions, while Küenle & Schurath (2003) study the average-reward optimality equation and stationary ε-optimal strategies under more general structural conditions.

On the learning side, Wei et al. (2017) study online average-reward zero-sum stochastic games. Their UCSG algorithm has regret guarantees governed by a game diameter, and an optimistic self-play construction yields guarantees for learning stationary maximin policies. Our setting combines strategic interaction with a separate worst-case transition evaluation for each player. These evaluations can use different adverse kernels, so robust payoffs can be general-sum even when the nominal rewards are zero-sum. The relevant structural control must therefore apply to every player's response problem, uniformly over stationary opponents.

A.2 AVERAGE-REWARD MDP LEARNING AND DISCOUNTED REDUCTIONS

In ordinary average-reward MDPs, diameter, mixing time, and bias span capture different aspects of long-run sensitivity. Jaksch et al. (2010) develop the diameter-based UCRL2 regret analysis for online learning, while Bartlett & Tewari (2009) use bias-span regularization for weakly communicating MDPs. These online formulations include the cost of reaching informative states. A generative model,

as assumed here, permits direct queries at each state–action row. Under this access model, Wang et al. (2024) obtain sample complexity linear in a uniform mixing-time bound, up to logarithmic factors, for uniformly ergodic MDPs. Zurek & Chen (2024b) establish span-based optimal sample complexity for weakly communicating models through discounted reduction and analyze general multichain models using an additional transient-time parameter. Zurek & Chen (2024a) further study direct planning in empirical average-reward models, with optimal diameter- and mixing-based guarantees and separate span-based upper and lower bounds.

Discounted approximation also has a planning interpretation. Grand-Clément & Petrik (2023) introduce a Blackwell discount factor above which every discounted-optimal policy of a finite MDP is also Blackwell- and average-optimal, and bound this threshold using the encoding size of the instance. Our reduction addresses approximate equilibria through an explicit bias-span bound on the Nash gap. It quantifies how discounted planning error and the choice of discount factor affect average-reward accuracy. Moreover, the span bound must hold uniformly over opponents because the empirical equilibrium determines the response problem to be controlled. As in the single-agent literature, a span parameter should be distinguished from a worst-case mixing or hitting time: the former depends on rewards as well as transition structure.

A.3 Robust dynamic programming and uncertainty models

Iyengar (2005); Nilim & El Ghaoui (2005) establish robust dynamic programming for uncertain transition laws under rectangularity conditions. Their results provide the discounted Bellman contraction and fixed-point theory used in our response problems. Wiesemann et al. (2013) develop robust MDP formulations with statistically motivated uncertainty sets and tractable optimization representations. A related distributionally robust formulation places ambiguity on probability distributions over uncertain model parameters (Xu & Mannor, 2010). In our formulation, nature directly chooses transition rows from prescribed sets around a nominal kernel; the nominal kernel is the object estimated from samples. The ambiguity radii describe the deployment uncertainty and are retained when forming the empirical game.

Rectangularity specifies which transition choices nature may combine. With uncertainty coupling different primitive rows, independently minimizing each row can assemble a kernel outside the original ambiguity set. The resulting Bellman operator evaluates a rectangular relaxation rather than, in general, the coupled model. Grand-Clément et al. (2024) connect robust-MDP tractability and dynamic programming to a simultaneous solvability property, clarifying the structural role of rectangularity. Our assumptions impose independence across primitive state–joint-action rows; after opponents are fixed, they yield the rectangular robust response model used in the proofs.

The geometry of the uncertainty set also affects planning. Ho et al. (2021) develop partial policy iteration and efficient Bellman updates for weighted L_1 ambiguity under both state and state–action rectangularity. These computational results complement the sample-complexity question studied here. We measure transition samples while allowing a discounted robust Nash planning oracle; efficient single-agent Bellman updates alone do not resolve general-sum equilibrium computation. For TV ambiguity, we also distinguish balls on the full simplex from balls restricted to a prescribed successor face. Only the former allow nature to move probability to an arbitrary minimum-value state, a distinction used in our span certificates.

A.4 Learning robust MDPs with discounted or finite-horizon rewards

Single-agent robust reinforcement learning provides the basic statistical problem of estimating a worst-case Bellman expectation. Panaganti & Kalathil (2022) give generative-model guarantees for robust MDPs with several divergence-based uncertainty sets. Yang et al. (2022) study both finite-sample behavior and asymptotic inference for robust values, treating state–action and state rectangularity. For episodic models, Xu et al. (2023) develop robust phased value learning for TV, chi-square, KL, and Wasserstein ambiguity, improving the state dependence of earlier sample bounds. Shi et al. (2023) analyze how the uncertainty level affects the sample complexity of discounted robust reinforcement learning. Their TV analysis also supplies a useful inverse-radius bound on discounted value span, which is closely related to our positive-radius certificate.

These works show why uncertainty geometry and the accuracy regime matter when comparing robust learning rates. Our average-reward bound uses the robust optimal-response span to quantify long-run sensitivity and exhibits the combination $H + \sigma H^2$. A further difficulty arises in games: even after the empirical kernel is fixed, different equilibria may induce different opponents and continuation values. Consequently, applying a single-agent guarantee separately to fixed opponents does not by itself control the response problems selected by an empirical Nash oracle.

A.5 Robust games with discounted and finite-horizon rewards

Kardeş et al. (2011) establish stationary robust equilibria in discounted stochastic games with model uncertainty and study their computation. Zhang et al. (2020a) develop robust Nash Q-learning with convergence guarantees under additional conditions, together with robust policy-gradient and actor–critic methods. Mcmahan et al. (2024) study planning with state-rectangular uncertainty through an equivalence between robust equilibria and equilibria of regularized games. They also identify computational hardness and uncertainty structures admitting efficient planning. These results motivate separating equilibrium computation from the transition-sample complexity measured in our oracle formulation.

Finite-horizon robust games have several distinct learning formulations. Ma et al. (2023) introduce a robust correlated-equilibrium concept and a decentralized robust V-learning algorithm with polynomial episode complexity. Shi et al. (2024b) study player-specific transition uncertainty and establish generative-model guarantees for robust Nash, correlated, and coarse correlated equilibria, together with lower bounds. Shi et al. (2024a) introduce uncertainty sets that depend on opponents' behavior and obtain robust coarse correlated equilibria with polynomial dependence on the number of players. Jiao & Li (2024) develop robust Q-FTRL with minimax guarantees for coarse correlated equilibria under a finite-horizon formulation with contamination ambiguity. Farhat et al. (2026) study learning robust games through online interaction, including TV and KL uncertainty, thereby addressing access through trajectories rather than independent queries to arbitrary rows. The equilibrium concept, ambiguity model, and data-access assumptions are therefore essential to comparisons of these guarantees.

Finite-horizon backward induction separates continuation values across stages. With independent samples for each stage, continuations computed from later-stage samples are independent of current-stage transition samples. Our stationary average-reward problem couples the opponents and continuation values through the same empirical kernel. Its analysis must control this dependence while retaining the long-run sensitivity measured by the response span.

A.6 Robust average-reward control: structure, planning, and learning

Fixing the opponents produces a robust MDP, making robust average-reward control the closest single-agent comparison. Wang et al. (2023a) develop a robust average-reward Bellman characterization and a vanishing-discount approach under compact state–action rectangular uncertainty and an all-policy, all-kernel unichain assumption. They also give robust relative value iteration under an additional recurrence condition. Wang & Si (2025) study constant-gain robust Bellman equations, their relation to stationary and history-dependent policies, and one-sided weak communication conditions. Our game-level analysis uses this response structure and additionally requires uniform control over opponents.

There are also complementary planning approaches. Chatterjee et al. (2023) reduce long-run average-reward polytopic robust MDPs to finite turn-based stochastic games and develop robust polytopic policy iteration. Here the second side represents nature, whereas our model also includes several strategic players, each with its own worst-case evaluation. Sun et al. (2024) develop robust policy mirror descent with global convergence guarantees for ergodic robust average-cost MDPs with general uncertainty sets. Wang et al. (2025) study projected policy gradient and worst-case transition evaluation for ergodic tabular robust average-reward MDPs with general ambiguity sets. These optimization results address uncertainty beyond the rowwise model considered here; their planning guarantees do not supply the uniform empirical-equilibrium control needed for our statistical theorem.

For learning, Chen et al. (2025) give finite-sample guarantees under nominal uniform ergodicity and sufficiently small KL or f_k-divergence ambiguity sets. Model-free, policy-evaluation, and discounted-

reduction approaches are studied under their respective recurrence and uncertainty assumptions in Wang et al. (2023b); Roch et al. (2025; 2026); Xu et al. (2025a;b). More recently, Yang et al. (2026) study TV ambiguity under a unichain assumption for every stationary policy and admissible kernel. Their bounds distinguish nominal and robust optimal bias spans and identify an additional term proportional to the ambiguity radius times the squared robust span in the low-tolerance regime. Our $H + \sigma H^2$ dependence reflects this robust-control sensitivity. The game analysis additionally handles opponents and equilibria chosen from the transition data and allows some multichain models through finite H_{R}.

A.7 Robust average-reward Markov games

The closest game-theoretic work is Roch & Wang (2026). They formulate average-reward distributionally robust Markov games and establish stationary robust Nash equilibrium existence under irreducibility, with a further existence result under fixed-opponent weak communication and uniqueness of the optimal robust bias up to additive constants. They also study discounted approximation and propose Robust Nash-Iteration and robust TD Descent. Their Nash-iteration convergence result uses additional structural and equilibrium-selection conditions; the TD-descent analysis establishes convergence to stationary points of a smoothed residual objective.

Our analysis centers on the uniform minimum optimal-response bias span H_{R}. Minimizing over Bellman solutions accommodates nonunique biases, while taking the supremum over stationary opponents supplies the uniformity needed at empirical equilibrium outputs. Under finite H_{R}, Theorem 3.3 establishes stationary statewise robust equilibrium existence and identifies every vanishing-discount equilibrium cluster point as an average-reward equilibrium. The explicit Nash-gap reduction in Theorem 3.4 determines the discounted accuracy sufficient for learning. This assumption requires a constant optimal response gain for each fixed opponent profile; it includes some multichain models but does not cover arbitrary state-dependent optimal-response gains. The resulting quantitative framework connects equilibrium existence to the statistical upper bound and the lower-bound constructions.

A.8 Nonlinear ergodicity and quantitative span certificates

Robust response operators are monotone and additively homogeneous, placing them within the theory of nonlinear Shapley operators. Akian et al. (2015) characterize reward-uniform ergodicity using the recession, or payment-free, operator and invariant-set conditions. Gaubert & Qu (2013) study Dobrushin-type contraction in Hopf's oscillation seminorm, which becomes the span seminorm on finite-state value vectors. Classical hitting-time comparisons also bound bias spans (Bartlett & Tewari, 2009). These results explain why communication, common transition mass, and reachability can control long-run values.

Our certificates specialize these mechanisms to robust responses and track the constants uniformly over opponents. The qualitative positive-radius full-simplex TV condition is already covered by the robust-control criteria of Wang & Si (2025), and inverse-radius span control is also present in discounted robust learning (Shi et al., 2023). We make their game-level implications explicit and study heterogeneous radii, restricted successor faces, and conditions that allow some zero-radius rows. The distinction between reward-uniform and reward-dependent conditions is useful here: communication can certify finite span for broad reward classes, while finite H_{R} for a particular reward need not imply weak communication.

A.9 Empirical equilibrium selection and statistical tightness

The dependence between an empirical model and its optimized value is central to model-based learning. Agarwal et al. (2020) use an absorbing-MDP leave-one-out construction to decouple discounted optimal values from individual transition samples. In discounted zero-sum games, Zhang et al. (2023) develop an absorbing-game analysis for learning the Nash value and establish policy guarantees using a smooth planning oracle. General-sum empirical games can instead have several equilibrium value branches, with a oracle choosing a branch using the same samples that enter its continuation value.

We address this dependence through one concentration event uniform over continuation vectors and a deterministic perturbation argument along a path of nominal centers. The resulting bound holds

simultaneously for every exact empirical discounted Nash oracle and also permits the approximate-oracle tolerance specified in Theorem 5.1. The lower bounds distinguish the basic span and ambiguity cost in an adaptive control embedding from the additional successor-support cost of balanced guarantees covering every exact empirical oracle. On the two-player, two-action class, the latter matches the upper bound jointly in $S, d_{\max}, H, \sigma$, and ε, up to logarithmic factors.

A guarantee covering every oracle is a different objective from designing one statistically favorable oracle. Our construction includes a oracle that avoids the hard equilibrium branch. For shared-reward, shared-ambiguity teams satisfying the all-policy, all-kernel unichain condition, centralized joint-action planning can also use the robust-MDP results of Yang et al. (2026), as detailed in Appendix I. Thus the support factor in our matching result concerns uniform guarantees over empirical oracles; the best achievable rate for a designed oracle in general-sum games with player-specific ambiguity remains a separate question.

B Model, information structure, and learning objective

Notation. For a positive integer m, let $[m] := \{1, \dots, m\}$. For a finite set $\mathcal{X}$, let $\Delta(\mathcal{X})$ denote the probability simplex on $\mathcal{X}$. We identify distributions and real-valued functions on a finite set with vectors indexed by that set. Write $\mathbf{1}$ for the all-ones vector, I for the identity matrix, and δ_x for the point mass at x, with dimensions determined by context. For a vector v, define

$$\|v\|_1 := \sum_x |v(x)|, \qquad \|v\|_\infty := \max_x |v(x)|, \qquad \mathrm{sp}(v) := \max_x v(x) - \min_x v(x).$$

For a probability vector p, write $\mathrm{supp}(p) := \{x : p(x) > 0\}$ and $p^\top v := \sum_x p(x)v(x)$. The variance notation used in the learning analysis is

$$\mathrm{Var}_p(v) := p^\top v^2 - (p^\top v)^2, \qquad \mathrm{Var}_Q(v) := Qv^2 - (Qv)^2$$

for a stochastic matrix Q; thus the coordinate of $\mathrm{Var}_Q(v)$ at s is the variance under row $Q(s, \cdot)$. Vector inequalities, squares, and square roots are interpreted coordinatewise. We use $\mathbb{P}$ and $\mathbb{E}$ for probability and expectation, $\mathbf{1}\{E\}$ for the indicator of an event E, and $(z)_+ := \max\{z, 0\}$ for the positive part of a scalar. The notation $\mathrm{KL}(p\|q)$ denotes Kullback–Leibler divergence, and $\mathrm{Ber}(u)$ denotes a Bernoulli distribution with mean u. The notation $\widetilde{O}$, $\widetilde{\Omega}$, and $\widetilde{\Theta}$ suppresses logarithmic factors.

B.1 Finite robust Markov game

We first review the formulations.

Let $\mathcal{N} = \{1, \dots, n\}$ be the players, $\mathcal{S}$ a finite state space of cardinality $S = |\mathcal{S}|$, and $\mathcal{A}_i$ a finite action set of cardinality $A_i = |\mathcal{A}_i|$. Write

$$\mathcal{A} \triangleq \prod_{i=1}^n \mathcal{A}_i, \qquad A_\times \triangleq |\mathcal{A}| = \prod_{i=1}^n A_i.$$

A stationary policy of player i is $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$. The product policy space is

$$\Pi_i = \prod_{s \in \mathcal{S}} \Delta(\mathcal{A}_i), \qquad \Pi = \prod_{i=1}^n \Pi_i.$$

For $\pi \in \Pi$, the joint action distribution at s is

$$\pi(\boldsymbol{a} \mid s) = \prod_{i=1}^n \pi_i(a_i \mid s).$$

Player i has known reward $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

The nominal transition kernel is

$$P^0 = \{P^0_{s,\boldsymbol{a}} \in \Delta(\mathcal{S}) : (s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}\}.$$

Player i evaluates performance under a player-specific state-joint-action rectangular total-variation ambiguity set

$$\mathcal{P}_i(P^0) = \prod_{(s,\boldsymbol{a})} \mathcal{P}_{i,s,\boldsymbol{a}}(P^0_{s,\boldsymbol{a}}),$$

where

$$\mathcal{P}_{i,s,\boldsymbol{a}}(q) = \{p \in \Delta(\mathcal{S}) : \mathrm{TV}(p,q) \le \sigma_{i,s,\boldsymbol{a}}\}, \qquad 0 \le \sigma_{i,s,\boldsymbol{a}} \le \sigma \le \frac{1}{2}.$$

The total-variation convention is $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$.

Nature is player specific: the kernel minimizing player i's payoff need not minimize another player's payoff. Nature observes the realized joint action before selecting its transition row. The history $\mathcal{F}_t$ is taken immediately before the actions at time t. Conditional on this history and the current state, the stationary opponents draw their actions independently according to θ and independently of the responding player's current randomization. An admissible history-dependent nature strategy may depend on this history and on the realized joint action, but its selected row must belong to the corresponding primitive ambiguity set. This information order is used in every conditional Bellman inequality below.

The existence, span, and deterministic reduction results require only nonempty compact local ambiguity sets and rectangularity across primitive rows. The total-variation form is needed for the statistical results.

B.2 Robust discounted values

Fix player i, opponents' stationary profile $\theta \in \Pi_{-i}$, and discount $\gamma \in (0,1)$. For $q \in \Delta(\mathcal{S})$ and $v \in \mathbb{R}^S$, define

$$\phi_{i,s,\boldsymbol{a}}(v;q) \triangleq \min_{p \in \mathcal{P}_{i,s,\boldsymbol{a}}(q)} p^\top v.$$

The robust discounted optimal-response operator is

$$(T^{\theta,Q}_{i,\gamma} v)(s) \triangleq \max_{a_i \in \mathcal{A}_i} \sum_{\boldsymbol{a}_{-i}} \theta(\boldsymbol{a}_{-i} \mid s) \left[r_i(s, a_i, \boldsymbol{a}_{-i}) + \gamma \phi_{i,s,(a_i,\boldsymbol{a}_{-i})}(v; Q_{s,(a_i,\boldsymbol{a}_{-i})}) \right]. \tag{12}$$

Here Q is a nominal center, which may be the population center, the empirical center, or a center on the segment between them. By standard rectangular robust discounted dynamic programming (Iyengar, 2005; Kardeş et al., 2011), the operator is monotone and a γ-contraction in the supremum norm. Its unique fixed point is denoted

$$V^{\star,\theta,Q}_{i,\gamma}.$$

For a stationary response $\mu_i \in \Pi_i$, replace the maximum in (12) by averaging over $\mu_i(\cdot \mid s)$; the unique fixed point is $V^{\mu_i,\theta,Q}_{i,\gamma}$. For a profile π, write

$$V^{\pi,Q}_{i,\gamma} \triangleq V^{\pi_i,\pi_{-i},Q}_{i,\gamma}.$$

Unless a center is displayed, $Q = P^0$.

The statewise discounted robust Nash gap is

$$\mathrm{Gap}^Q_\gamma(\pi) \triangleq \max_{i \in \mathcal{N}} \left\| V^{\star,\pi_{-i},Q}_{i,\gamma} - V^{\pi,Q}_{i,\gamma} \right\|_\infty. \tag{13}$$

The difference inside the norm is componentwise nonnegative. A profile is an e_γ-discounted robust Nash equilibrium when $\mathrm{Gap}^{P^0}_\gamma(\pi) \le e_\gamma$.

B.3 Robust average rewards

For a possibly history-dependent response ν_i, stationary opponents θ, an admissible nature strategy τ, and initial state s, write

$$J_{i,T}(\nu_i, \theta, \tau; s) \triangleq \frac{1}{T} \mathbb{E}^{\nu_i,\theta,\tau}_s \left[\sum_{t=0}^{T-1} r_i(S_t, \boldsymbol{A}_t) \right].$$

A stationary kernel $P \in \mathcal{P}_i(P^0)$ is a special case of τ. Let $\mathfrak{T}_i$ be the set of all admissible history-dependent nature strategies. Define

$$\underline{g}_i^{(\nu_i,\theta)}(s) \triangleq \inf_{P\in\mathcal{P}_i(P^0)} \liminf_{T\to\infty} J_{i,T}(\nu_i,\theta,P;s),$$
$$\underline{g}_{i,\mathrm{dyn}}^{(\nu_i,\theta)}(s) \triangleq \inf_{\tau\in\mathfrak{T}_i} \liminf_{T\to\infty} J_{i,T}(\nu_i,\theta,\tau;s).$$

The main paper uses the first convention. The second records the stronger nature model explicitly. If all players and the kernel are stationary, the expected Cesaro average has a limit by finite-state Markov-chain theory, so the liminf in the first definition may be replaced by a limit for such profiles. No limit is assumed for general history-dependent strategies.

For stationary responses, the optimal robust gain is

$$\rho_i^\star(\theta,s) \triangleq \sup_{\mu_i\in\Pi_i} \underline{g}_i^{(\mu_i,\theta)}(s).$$

For a stationary product profile π, write $\underline{g}_i^\pi = \underline{g}_i^{(\pi_i,\pi_{-i})}$ and define

$$\mathrm{Gap}_{\mathrm{av}}(\pi) \triangleq \max_{i\in\mathcal{N}} \max_{s\in\mathcal{S}} \left[\rho_i^\star(\pi_{-i},s) - \underline{g}_i^\pi(s)\right]. \tag{14}$$

A profile is an ε-robust average-reward Nash equilibrium if $\mathrm{Gap}_{\mathrm{av}}(\pi) \le \varepsilon$. Under Assumption C.2, Proposition C.6 shows that the optimal gain is independent of s, that allowing history-dependent responses does not increase it, and that it is the same under either nature convention. We then write $\rho_i^\star(\theta)$. The fixed-policy lower bounds below hold also for $\underline{g}_{i,\mathrm{dyn}}^\pi$. Thus every stated average-reward guarantee also holds when the played policy is evaluated against history-dependent nature.

B.4 Generative model and oracle-uniform plug-in interface

For every primitive state–joint-action row $x = (s, \boldsymbol{a})$, the balanced experiment supplies N independent samples

$$S'_{x,1},\dots,S'_{x,N} \overset{\mathrm{iid}}{\sim} P_x^0.$$

The empirical center is

$$\widehat{P}_x^0(s') = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}\{S'_{x,k} = s'\}.$$

Samples from distinct primitive rows are also independent. The total number of transition samples is

$$N_{\mathrm{tot}} = SA_\times N.$$

The learner knows rewards and ambiguity radii.

For a discount factor γ, let $\mathcal{E}_\gamma(\widehat{P}^0)$ denote the set of stationary statewise robust Nash equilibria of the empirical discounted game. An exact Nash oracle is a measurable oracle

$$\mathfrak{O}_\gamma(\widehat{P}^0) \in \mathcal{E}_\gamma(\widehat{P}^0).$$

The learning theorem is *oracle uniform*: it holds for every measurable exact oracle $\mathfrak{O}_\gamma$. This is stronger than fixing one statistically designed selection rule. The oracle is a computational abstraction and is not charged to the transition-sample budget.

The lower-bound discussion also considers the unrestricted adaptive generative-model experiment, in which the learner chooses the next primitive row after observing its previous queries and samples. This stronger sampling model is used only to interpret the information-theoretic lower scale; Algorithm 1 uses the balanced experiment above.

C Uniform robust optimal-response span

This section defines the main analytic assumption. It includes constant-gain Bellman solvability and a uniform minimum-span bound in one quantity.

C.1 CONSTANT-GAIN RESPONSE EQUATIONS

Fix player i and stationary opponents $\theta \in \Pi_{-i}$. Define the average-reward robust optimal-response operator

$$(T_i^\theta h)(s) \triangleq \max_{a_i \in \mathcal{A}_i} \sum_{\boldsymbol{a}_{-i}} \theta(\boldsymbol{a}_{-i} \mid s) \left[r_i(s, a_i, \boldsymbol{a}_{-i}) + \phi_{i,s,(a_i,\boldsymbol{a}_{-i})}(h; P^0_{s,(a_i,\boldsymbol{a}_{-i})}) \right]. \tag{15}$$

For later use, define the opponent-averaged reward and robust continuation functional by

$$\bar{r}_i^\theta(s, a_i) \triangleq \sum_{\boldsymbol{a}_{-i}} \theta(\boldsymbol{a}_{-i} \mid s) r_i(s, a_i, \boldsymbol{a}_{-i}), \tag{16}$$

$$\bar{\phi}_i^\theta(v; s, a_i) \triangleq \sum_{\boldsymbol{a}_{-i}} \theta(\boldsymbol{a}_{-i} \mid s) \phi_{i,s,(a_i,\boldsymbol{a}_{-i})} \left(v; P^0_{s,(a_i,\boldsymbol{a}_{-i})} \right). \tag{17}$$

Then

$$(T_i^\theta h)(s) = \max_{a_i} \left\{ \bar{r}_i^\theta(s, a_i) + \bar{\phi}_i^\theta(h; s, a_i) \right\}.$$

A constant-gain Bellman solution is a pair $(\rho, h) \in \mathbb{R} \times \mathbb{R}^S$ satisfying

$$\rho \mathbf{1} + h = T_i^\theta h. \tag{18}$$

Let

$$\mathfrak{B}_i(\theta) \triangleq \left\{ (\rho, h) \in \mathbb{R} \times \mathbb{R}^S : \rho \mathbf{1} + h = T_i^\theta h \right\}. \tag{19}$$

We adopt the convention $\inf \varnothing = +\infty$.

Definition C.1. *For player i and opponents θ, define*

$$H_i(\theta) \triangleq \inf_{(\rho,h) \in \mathfrak{B}_i(\theta)} \max\{1, \mathrm{sp}(h)\}. \tag{20}$$

The game-level robust optimal-response span is

$$H_{\mathrm{R}} \triangleq \max_{i \in \mathcal{N}} \sup_{\theta \in \Pi_{-i}} H_i(\theta). \tag{21}$$

Assumption C.2 (Uniform bounded optimal-response span)**.** *The game satisfies*

$$H_{\mathrm{R}} < \infty. \tag{22}$$

Write $H \geq H_{\mathrm{R}}$ for any finite upper bound. Existence and the deterministic comparison results do not require this bound to be known. A supplied bound is needed only when choosing the discount or sample size algorithmically.

Because $\inf \varnothing = +\infty$, Assumption C.2 guarantees that every fixed-opponent response problem has at least one constant-gain Bellman solution. It does not require uniqueness of the bias or of the maximizing policy.

For applications of single-agent robust dynamic programming, define the induced response row sets

$$\mathcal{Q}^{i,\theta}_{s,a_i} = \left\{ \sum_{\boldsymbol{a}_{-i}} \theta(\boldsymbol{a}_{-i} \mid s) p_{\boldsymbol{a}_{-i}} : p_{\boldsymbol{a}_{-i}} \in \mathcal{P}_{i,s,(a_i,\boldsymbol{a}_{-i})} \right\}.$$

They are nonempty compact sets of probability vectors, being continuous images of finite products of the primitive sets. Primitive rectangularity gives

$$\min_{q \in \mathcal{Q}^{i,\theta}_{s,a_i}} q^\top v = \bar{\phi}_i^\theta(v; s, a_i),$$

and permits independent selections across (s, a_i). Every induced row selection lifts to admissible primitive rows.

Lemma C.3. *For every local ambiguity set and vectors $u, v \in \mathbb{R}^S$:*

(i) if $u \leq v$, then $\phi(u; q) \leq \phi(v; q)$;

(ii) $\phi(v + c\mathbf{1}; q) = \phi(v; q) + c$ *for every* $c \in \mathbb{R}$*;*

(iii) $|\phi(u; q) - \phi(v; q)| \le \|u - v\|_\infty$.

Proof. Every feasible p is a probability vector, so $u \le v$ implies $p^\top u \le p^\top v$, and $p^\top(v + c\mathbf{1}) = p^\top v + c$. Minimization proves (i)–(ii). Apply them to $v - \|u - v\|_\infty \mathbf{1} \le u \le v + \|u - v\|_\infty \mathbf{1}$ to obtain (iii). □

Proposition C.4. *For fixed i, θ, and $\gamma \in (0, 1)$, the optimal-response operator $T_{i,\gamma}^{\theta,P^0}$ and each stationary-policy evaluation operator are monotone γ-contractions in $\|\cdot\|_\infty$. Their fixed points equal the corresponding robust discounted values under both stationary-kernel nature and history-dependent nature. The optimal fixed point is optimal also against all history-dependent controller responses. A stationary policy is optimal from every state if and only if it is supported on the maximizing actions in the optimal fixed-point equation.*

Proof. The contraction, fixed-point characterization, and stationary optimality are standard rectangular robust discounted dynamic programming (Iyengar, 2005). The required hypotheses hold for the induced response model: states and controller actions are finite, rewards $\bar{r}_i^\theta$ are bounded, and the nonempty compact row sets $\mathcal{Q}_{s,a_i}^{i,\theta}$ can be selected independently. For policy evaluation, average additionally over the fixed own-action distribution. Replacing any local set by its compact convex hull leaves all linear minima unchanged. Hence the convex-set formulation of the standard result also applies, and minimizing rows can still be chosen in the original sets.

The stationary-nature guarantee against history-dependent responses uses a single primitive row minimizer P^V at the optimal fixed point V. Every own action then has discounted score at most $V(s)$ under P^V. Discounted conditional telescoping bounds every response's return by $V(s)$, since the terminal term has absolute value at most $\gamma^T \|V\|_\infty \to 0$. A maximizing stationary response satisfies the reverse inequality against every admissible nature strategy. Finally, a stationary response is optimal in every state exactly when its evaluation operator fixes V. An average of action scores bounded above by $V(s)$ equals $V(s)$ exactly when all positive-weight actions are maximizers. This proves the support characterization. □

C.2 Bellman verification and minimum-span solutions

The following verification argument makes the information structure explicit. The same primitive minimizing kernel controls every unilateral response, including responses that depend on the past. This is the feature supplied by state–joint-action rectangularity.

Lemma C.5 (One-sided Bellman verification)**.** *Fix i, stationary opponents θ, and $w \in \mathbb{R}^S$. For a stationary response μ_i, define*

$$(F^{\mu_i} w)(s) := \sum_{a_i} \mu_i(a_i \mid s)\big[\bar{r}_i^\theta(s, a_i) + \bar{\phi}_i^\theta(w; s, a_i)\big].$$

If $F^{\mu_i} w \ge a\mathbf{1} + w$, then both $\underline{g}_i^{(\mu_i,\theta)}(s)$ and $\underline{g}_{i,\mathrm{dyn}}^{(\mu_i,\theta)}(s)$ are at least a. If $F^{\mu_i} w \le b\mathbf{1} + w$, then both payoffs are at most b. If $T_i^\theta w \le b\mathbf{1} + w$, the upper bound holds for every history-dependent randomized response. These statements require no constant-gain Bellman solution.

Proof. For the lower bound, fix any admissible nature strategy τ and let $\mathcal{F}_t$ be the history immediately before the actions at time t. Write $R_t = r_i(S_t, \boldsymbol{A}_t)$. Each selected primitive row has expectation of w at least its rowwise minimum. Averaging over the independent current actions therefore gives

$$\mathbb{E}^{\mu_i,\theta,\tau}[R_t + w(S_{t+1}) \mid \mathcal{F}_t] \ge (F^{\mu_i} w)(S_t) \ge a + w(S_t).$$

Summing expectations for $t = 0, \dots, T - 1$ yields

$$J_{i,T}(\mu_i, \theta, \tau; s) \ge a + \frac{w(s) - \mathbb{E}_s^{\mu_i,\theta,\tau} w(S_T)}{T} \ge a - \frac{\mathrm{sp}(w)}{T}.$$

The estimate holds for every τ and hence for stationary nature as well. Taking the liminf and then the infimum proves both lower bounds.

For the upper bound, choose, at every primitive row, a minimizer

$$P^w_{s,\boldsymbol{a}} \in \operatorname*{arg\,min}_{p \in \mathcal{P}_{i,s,\boldsymbol{a}}} p^\top w.$$

Compactness ensures attainment and rectangularity makes P^w an admissible stationary kernel. Under P^w, the first conditional inequality above becomes equality, so $F^{\mu_i} w \le b\mathbf{1} + w$ gives $J_{i,T}(\mu_i, \theta, P^w; s) \le b + \operatorname{sp}(w)/T$. If instead $T_i^\theta w \le b\mathbf{1} + w$, every own-action score obeys the same upper bound. Averaging it under the conditional action distribution of any history-dependent response gives $J_{i,T}(\nu_i, \theta, P^w; s) \le b + \operatorname{sp}(w)/T$. Taking the limsup proves the asserted upper bounds under stationary nature. They also hold under history-dependent nature because P^w is an admissible choice in that larger class. □

Proposition C.6 (Constant-gain optimality). *If $(\rho, h) \in \mathfrak{B}_i(\theta)$, then, for every initial state s,*

$$\sup_{\nu_i} \underline{g}_i^{(\nu_i,\theta)}(s) = \sup_{\nu_i} \underline{g}_{i,\mathrm{dyn}}^{(\nu_i,\theta)}(s) = \rho,$$

where both suprema allow all history-dependent randomized responses. Every stationary response supported on maximizing actions in (18) *attains ρ under both nature conventions. The statements remain true if liminf is replaced by limsup in the payoffs.*

Proof. This is the rectangular specialization of robust Bellman verification (Wang & Si, 2025, Theorems 2–3). In the present model it follows directly from Lemma C.5: $T_i^\theta h = \rho\mathbf{1} + h$ bounds every response above by ρ, whereas a stationary maximizing response μ_i satisfies $F^{\mu_i} h = \rho\mathbf{1} + h$ and guarantees at least ρ. The proof of that lemma bounds the finite-horizon averages from below and above with errors $\operatorname{sp}(h)/T$, which also establishes the limsup assertion. No exchange of an infimum with an infinite-horizon limit is used. □

Consequently, every solution in $\mathfrak{B}_i(\theta)$ has the same gain, which equals $\rho_i^\star(\theta, s)$ for all s. We denote it by $\rho_i^\star(\theta)$ and reserve the calligraphic symbol

$$\mathcal{H}_i(\theta) := \{h : (\rho_i^\star(\theta), h) \in \mathfrak{B}_i(\theta)\}$$

for the bias-solution set. Thus $H_i(\theta)$ is always a scalar.

Lemma C.7 (Attainment of the minimum span). *If $H_i(\theta) < \infty$, there is $h_i^{\min}(\theta) \in \mathcal{H}_i(\theta)$ whose span is minimal, and*

$$\max\{1, \operatorname{sp}\big(h_i^{\min}(\theta)\big)\} = H_i(\theta).$$

Proof. Use the compact-sublevel argument of (Yang et al., 2026, Proposition 4). The solution set is nonempty here by definition of $H_i(\theta)$. Normalize every bias by $\min h = 0$ and take a sequence whose spans decrease to their infimum. Additive homogeneity preserves the equation, and the one-Lipschitz property of T_i^θ makes the normalized solution set closed. The sequence lies in some cube $[0, C]^S$, so a convergent subsequence has a limit in that solution set. Continuity of the span gives attainment. Applying $\max\{1, \cdot\}$ proves the displayed identity. □

Remark C.8 (Scope of the span assumption). *Assumption C.2 requires a constant optimal robust gain for every stationary opponent profile. It can nevertheless admit multichain models and does not imply weak communication. For example, two disconnected absorbing states, one action per player, singleton ambiguity sets, and the same reward at both states have a zero bias and $H_{\mathrm{R}} = 1$. Neither state is transient and the states cannot communicate. Models whose optimal robust response gain depends on the initial state are outside the assumption.*

C.3 Equivalent discounted-span formulation

Define

$$D_{\mathrm{R}} \triangleq \max\left\{1, \max_{i \in \mathcal{N}} \sup_{\theta \in \Pi_{-i}} \sup_{\gamma \in (0,1)} \operatorname{sp}\Big(V_{i,\gamma}^{\star,\theta}\Big)\right\}. \tag{23}$$

Theorem C.9. *For the finite robust Markov game,*

$$H_{\mathrm{R}} \le D_{\mathrm{R}} \le 2H_{\mathrm{R}} \tag{24}$$

in the extended nonnegative reals. Consequently,

$$H_{\mathrm{R}} < \infty \quad \Longleftrightarrow \quad D_{\mathrm{R}} < \infty.$$

Proof. We apply (Wang & Si, 2025, Theorem 4 and Appendix B) to each response problem. The induced row sets above give precisely its finite-state robust MDP formulation, with rewards $\bar{r}_i^{\theta}$ and controller action distributions $\Delta(\mathcal{A}_i)$. No communication or bias-uniqueness assumption is required.

Write $d_i(\theta) = \sup_{\gamma\in(0,1)} \operatorname{sp}\left(V_{i,\gamma}^{\star,\theta}\right)$. If $d_i(\theta) < \infty$, the normalized vanishing-discount construction in that proof produces a Bellman solution with span at most $d_i(\theta)$: its discounted biases lie in $[0, d_i(\theta)]^S$, and the scaled minimum values lie in $[0, 1]$. Hence $H_i(\theta) \le \max\{1, d_i(\theta)\}$.

Conversely, choose a minimum-span solution (ρ, h) with $\min h = 0$ and $B = \operatorname{sp}(h)$ whenever $H_i(\theta) < \infty$. The subsolution–supersolution comparison in the same Appendix B, equation (B.2), gives

$$\frac{\rho}{1-\gamma}\mathbf{1} + h - B\mathbf{1} \le V_{i,\gamma}^{\star,\theta} \le \frac{\rho}{1-\gamma}\mathbf{1} + h. \tag{25}$$

Its hypotheses are exactly $\rho\mathbf{1} + h = T_i^{\theta} h$, $0 \le h \le B\mathbf{1}$, and monotonicity and contraction of the discounted operator. Taking spans gives $d_i(\theta) \le 2B \le 2H_i(\theta)$. Taking the required suprema and maxima proves (24); cases with an infinite right-hand side are automatic. □

Remark C.10. *A uniform response analysis must control the problem induced by the random opponents returned by the empirical Nash oracle. An assumption imposed only at population equilibria does not by itself give this control: the empirical opponents need not yet be population-equilibrium opponents. The supremum in* (21) *supplies the bound needed by the present proof. The following example shows that a population-equilibrium span bound cannot be substituted for this uniform response bound.*

Example C.11 (Bounded equilibrium span and large off-equilibrium span)**.** *Fix $\eta \in (0, 1/2)$. There are two states $0, 1$, no ambiguity, and two players. Player 1 has one action and reward $r_1(s) = s$. Player 2 has actions f and ℓ, with reward $r_2(s, f) = 1$ and $r_2(s, \ell) = 0$ at both states. Action f sends the next state to the uniform distribution on $\{0, 1\}$. Action ℓ switches to the other state with probability η and otherwise stays.*

For a stationary policy of player 2, let α_s be its probability of choosing f at state s. Its probability of switching states is

$$p_s = \frac{\alpha_s}{2} + (1 - \alpha_s)\eta > 0.$$

Consequently every such policy induces an irreducible two-state chain. Its invariant distribution d_α assigns positive mass to both states, and player 2's gain is $d_\alpha(0)\alpha_0 + d_\alpha(1)\alpha_1$. This gain equals one if and only if $\alpha_0 = \alpha_1 = 1$. Player 2 can guarantee one by choosing f at both states, so every stationary average-reward equilibrium uses this policy. Player 1 has no deviation.

The player 1 Bellman equations, normalized by $h(0) = 0$, are

$$\rho = p_0 h(1), \qquad \rho + h(1) = 1 + (1 - p_1)h(1).$$

Subtracting the first equation from the second gives $h(1) = 1/(p_0 + p_1)$. Thus the minimum bias span is one at equilibrium and $1/(2\eta)$ when player 2 always chooses ℓ. Player 2's optimal-response equation has the solution $(\rho, h) = (1, 0)$, so its span parameter is one. The equilibrium-local game parameter is therefore one, whereas $H_{\mathrm{R}} = 1/(2\eta)$ diverges as $\eta \downarrow 0$.

D EQUILIBRIUM EXISTENCE

This section proves equilibrium existence. It first establishes discounted robust equilibria, then verifies their vanishing-discount limits, and finally states the consequence of the bounded optimal-response span assumption.

D.1 Existence of discounted robust equilibria

For each player, define the discounted greedy-response correspondence

$$B_i^\gamma(\theta) \triangleq \left\{ \mu_i \in \Pi_i : \operatorname{supp} \mu_i(\cdot \mid s) \subseteq \arg\max_{a_i} \left[\bar{r}_i^\theta(s, a_i) + \gamma \bar{\phi}_i^\theta(V_{i,\gamma}^{\star,\theta}; s, a_i) \right] \ \forall s \right\}.$$

Lemma D.1. *For fixed i and γ, the map*

$$\theta \longmapsto V_{i,\gamma}^{\star,\theta}$$

is continuous on Π_{-i} in the supremum norm.

Proof. Let $\theta_k \to \theta$. Continuous dependence of a contraction's fixed point gives, for $V_k = V_{i,\gamma}^{\star,\theta_k}$ and $V = V_{i,\gamma}^{\star,\theta}$,

$$\|V_k - V\|_\infty \le \frac{\left\| T_{i,\gamma}^{\theta_k} V - T_{i,\gamma}^{\theta} V \right\|_\infty}{1-\gamma} \longrightarrow 0.$$

The numerator tends to zero because it consists of finite opponent averages of fixed action scores, followed by a finite maximum. □

Lemma D.2. *For every i and γ, B_i^γ has nonempty, compact, convex values and a closed graph. Consequently it is upper hemicontinuous.*

Proof. The action scores are continuous in θ by Lemma D.1 and the support-function Lipschitz bound. At each state, $B_i^\gamma(\theta)$ consists of the probability distributions maximizing the continuous linear objective obtained by averaging those scores. Berge's maximum theorem gives nonempty compact values and upper hemicontinuity on the fixed compact simplex. Linearity gives convex values. The finite product over states preserves these properties, and upper hemicontinuity with compact values gives a closed graph. □

Theorem D.3. *For every $\gamma \in (0,1)$, the finite discounted DR-MG admits a stationary profile π^γ such that*

$$\pi_i^\gamma \in B_i^\gamma(\pi_{-i}^\gamma) \qquad \forall i.$$

It is a robust discounted Nash equilibrium from every initial state, also against unilateral deviations that depend on the history.

Proof. This follows the discounted robust-game existence argument of (Kardeş et al., 2011). Its compactness and continuity conditions are verified here by Lemma D.2, separately for each player's ambiguity sets. Specifically, $B^\gamma(\pi) = \prod_i B_i^\gamma(\pi_{-i})$ is an upper hemicontinuous correspondence with nonempty compact convex values on the nonempty compact convex policy space Π. Kakutani's theorem gives a fixed point. Proposition C.4 identifies this fixed point as a statewise equilibrium even against history-dependent unilateral deviations. Thus allowing different ambiguity sets for different players requires no common adverse kernel. □

Measurable oracle outputs in the sampling experiment. For a fixed integer sample count N, each empirical primitive row has coordinates in $\{0, 1/N, \dots, 1\}$, so the set of possible empirical kernels is finite. Select one equilibrium, whose existence was just proved, for each possible empirical kernel. Every resulting map on this finite set is measurable. Thus the exact measurable oracle interface used in the learning theorem is nonempty without requiring uniqueness of an equilibrium.

D.2 Vanishing-discount limits and average-reward existence

The substantive limiting step must handle opponents that vary with the discount. We pass both the response-optimality equation and the selected policy's equality to the limit. This avoids requiring continuity of average payoffs or a uniquely chosen bias.

Theorem D.4. *Under Assumption C.2, finite state and action spaces, and nonempty compact state–joint-action rectangular ambiguity sets, a stationary statewise robust average-reward Nash equilibrium exists. For any sequence $\gamma_k \uparrow 1$ and stationary statewise robust discounted equilibria π^k, every cluster point of $\{\pi^k\}$ is such an equilibrium. The conclusion holds against history-dependent unilateral deviations under either nature convention, without bias uniqueness.*

Proof. Choose any sequence $\gamma_k \uparrow 1$. Theorem D.3 supplies a stationary discounted equilibrium π^k for every k, and compactness of Π supplies a cluster point. To verify any such point, restrict to a subsequence satisfying $\pi^k \to \pi^\star$. For each i, write $\theta_k = \pi^k_{-i}$, $\theta^\star = \pi^\star_{-i}$, and normalize

$$v_k = V^{\star,\theta_k}_{i,\gamma_k}, \qquad m_k = \min v_k, \qquad h_k = v_k - m_k\mathbf{1}, \qquad \beta_k = (1-\gamma_k)m_k.$$

Theorem C.9 gives $0 \le h_k \le 2H\mathbf{1}$, and bounded rewards give $0 \le \beta_k \le 1$. Because the number of players is finite, a further subsequence makes $h_k \to h_i$ and $\beta_k \to \beta_i$ simultaneously for every player. This extraction leaves the already fixed profile limit $\pi^\star$ unchanged.

Translation of the discounted Bellman equation gives

$$h_k(s) = \max_{a_i} \left\{ \bar{r}_i^{\theta_k}(s,a_i) - \beta_k + \gamma_k \bar{\phi}_i^{\theta_k}(h_k; s, a_i) \right\}. \tag{26}$$

Finite opponent averages and the one-Lipschitz support property make the scores jointly continuous in (θ, h, γ). Taking limits yields

$$\beta_i + h_i(s) = \max_{a_i} \left\{ \bar{r}_i^{\theta^\star}(s,a_i) + \bar{\phi}_i^{\theta^\star}(h_i; s, a_i) \right\}. \tag{27}$$

By Proposition C.4, averaging the scores in (26) against π_i^k leaves the maximum unchanged. Taking limits also gives

$$h_i(s) = \sum_{a_i} \pi_i^\star(a_i \mid s) \left[\bar{r}_i^{\theta^\star}(s,a_i) - \beta_i + \bar{\phi}_i^{\theta^\star}(h_i; s, a_i) \right]. \tag{28}$$

Every bracketed action score in (28) is at most $h_i(s)$ by (27). Their average equals $h_i(s)$, so every action receiving positive probability under $\pi_i^\star$ must attain the maximum. Proposition C.6 now gives, directly under both nature conventions,

$$\underline{g}_i^{\pi^\star}(s) \ge \beta_i \tag{29}$$

and, for every history-dependent response ν_i,

$$\underline{g}_i^{(\nu_i,\theta^\star)}(s) \le \beta_i. \tag{30}$$

Thus the limiting profile is a statewise robust average-reward equilibrium. The argument applies to every convergent equilibrium subsequence; it does not require continuity of average rewards or a unique bias selected as a function of opponents. □

E Deterministic discounted-to-average reduction

This section isolates the reduction theorem. It is independent of sampling and is the principal interface between average-reward theory and any discounted robust-game learner.

E.1 A fixed-policy Abelian sandwich

Lemma E.1. *Fix player i, opponents θ, a stationary response μ_i, and $\gamma \in (0,1)$. Let*

$$v = V^{\mu_i,\theta}_{i,\gamma}, \qquad m = \min_s v(s), \qquad M = \max_s v(s).$$

Then, for every initial state s,

$$(1-\gamma)m \le \underline{g}_i^{(\mu_i,\theta)}(s) \le (1-\gamma)M. \tag{31}$$

Both bounds remain valid against history-dependent rowwise choices by nature.

Proof. By the rectangular discounted policy-evaluation result in Proposition C.4,

$$v(s) = \sum_{a_i, \boldsymbol{a}_{-i}} \mu_i(a_i \mid s)\theta(\boldsymbol{a}_{-i} \mid s) \left[r_i(s, a_i, \boldsymbol{a}_{-i}) + \gamma \phi_{i,s,(a_i,\boldsymbol{a}_{-i})}(v; P^0_{s,(a_i,\boldsymbol{a}_{-i})}) \right]$$
$$\triangleq r^{\mu_i,\theta}(s) + \gamma \Phi_i^{\mu_i,\theta}(v)(s). \tag{32}$$

Every continuation is an expectation of v, so $m\mathbf{1} \le \Phi_i^{\mu_i,\theta}(v) \le M\mathbf{1}$. Consequently

$$(1-\gamma)m\mathbf{1} \le r^{\mu_i,\theta} + \Phi_i^{\mu_i,\theta}(v) - v \le (1-\gamma)M\mathbf{1}.$$

Apply Lemma C.5 with test vector v. It gives (31) under both nature conventions. No fixed-policy constant-gain solution or recurrence assumption is used. □

E.2 Optimal discounted values from a minimum-span bias

The vector comparison (25) immediately gives a one-sided estimate that is sharper than using only $\operatorname{sp}\left(V^{\star,\theta}_{i,\gamma}\right) \le 2H_i(\theta)$.

Corollary E.2. *Fix i and θ with $H_i(\theta) < \infty$. For every $\gamma \in (0,1)$,*

$$\rho_i^\star(\theta) \le (1-\gamma) \min_s V^{\star,\theta}_{i,\gamma}(s) + (1-\gamma)H_i(\theta). \tag{33}$$

Proof. For a minimum-span solution normalized by $\min h = 0$, the lower bound in (25) gives $(1-\gamma)\min V^{\star,\theta}_{i,\gamma} \ge \rho_i^\star(\theta) - (1-\gamma)\operatorname{sp}(h)$. Use $\operatorname{sp}(h) \le H_i(\theta)$ and rearrange. □

Theorem E.3. *Let $\pi \in \Pi$ be any stationary profile and let*

$$e_\gamma \triangleq \operatorname{Gap}^{P^0}_\gamma(\pi).$$

Under Assumption C.2,

$$\operatorname{Gap}_{\mathrm{av}}(\pi) \le (1-\gamma)(H_{\mathrm{R}} + e_\gamma). \tag{34}$$

More precisely, for every player i and state s,

$$\rho_i^\star(\pi_{-i}) - \underline{g}_i^\pi(s) \le (1-\gamma)\left(H_i(\pi_{-i}) + \left\| V^{\star,\pi_{-i}}_{i,\gamma} - V^\pi_{i,\gamma} \right\|_\infty \right).$$

The playerwise bound also holds with $\underline{g}_i^\pi(s)$ replaced by $\underline{g}^\pi_{i,\mathrm{dyn}}(s)$ and with arbitrary history-dependent unilateral deviations in the optimal response.

Proof. Fix i, put $\theta = \pi_{-i}$, and let $e_i = \left\| V^{\star,\theta}_{i,\gamma} - V^\pi_{i,\gamma} \right\|_\infty$. Corollary E.2 and Lemma E.1 imply

$$\rho_i^\star(\theta) - \underline{g}_i^\pi(s) \le (1-\gamma)\left[H_i(\theta) + \min V^{\star,\theta}_{i,\gamma} - \min V^\pi_{i,\gamma} \right]$$
$$\le (1-\gamma)\big(H_i(\theta) + e_i\big),$$

since $|\min u - \min v| \le \|u - v\|_\infty$. Maximizing over i, s proves (34). The strengthened version uses the same fixed-policy lower bound under history-dependent nature and Proposition C.6 for the optimal response. □

Corollary E.4. *Let $\varepsilon \in (0,1]$ and suppose $H \ge H_{\mathrm{R}}$ is known. Set*

$$\gamma = 1 - \frac{\varepsilon}{2H}. \tag{35}$$

Any profile satisfying $\operatorname{Gap}^{P^0}_\gamma(\pi) \le H$ is an ε-robust average-reward Nash equilibrium.

Proof. Because $H \ge H_{\mathrm{R}}$,

$$(1-\gamma)(H_{\mathrm{R}} + \operatorname{Gap}^{P^0}_\gamma(\pi)) \le \frac{\varepsilon}{2H}(H + H) = \varepsilon.$$

Applying Theorem E.3 then proves the claim. □

The reduction permits a coarse discounted tolerance. The discounted value itself has scale $(1-\gamma)^{-1} \asymp H/\varepsilon$, yet the solver only needs an additive value error of order H, not order ε. This discounted tolerance is the one used by the statistical analysis.

E.3 Additional deterministic certificates

Proposition E.5. *Fix i, stationary opponents θ, and $\gamma \in (0,1)$. Set*

$$V^* = V^{\star,\theta}_{i,\gamma}, \qquad h = V^* - \min_s V^*(s)\mathbf{1}, \qquad \beta = (1-\gamma)\min_s V^*(s).$$

For a stationary response μ_i, define

$$d(s) \triangleq h(s) - \sum_{a_i} \mu_i(a_i \mid s)\left[\bar{r}_i^{\theta}(s,a_i) - \beta + \gamma \bar{\phi}_i^{\theta}(h;s,a_i)\right],$$
$$\eta_i^{\gamma}(\mu_i;\theta) \triangleq \max_s d(s).$$

Then, from every initial state s,

$$\sup_{\nu_i} \underline{g}_i^{(\nu_i,\theta)}(s) - \underline{g}_i^{(\mu_i,\theta)}(s) \le \eta_i^{\gamma}(\mu_i;\theta) + (1-\gamma)\,\mathrm{sp}(V^*),$$

where the supremum permits all history-dependent randomized responses. The inequality also holds with both payoffs evaluated against history-dependent nature. Constant-gain Bellman solvability is not required.

Proof. Write $B = \mathrm{sp}(h) = \mathrm{sp}(V^*)$, $\eta = \eta_i^{\gamma}(\mu_i;\theta)$, and $\Phi^{\mu_i}(h) = \sum_{a_i} \mu_i(a_i \mid \cdot)\bar{\phi}_i^{\theta}(h;\cdot,a_i)$. The discounted optimality equation gives $0 \le d \le \eta\mathbf{1}$, while normalization gives $0 \le h \le B\mathbf{1}$. Hence the fixed-policy average-reward operator satisfies

$$F^{\mu_i}h - h = \beta\mathbf{1} - d + (1-\gamma)\Phi^{\mu_i}(h) \ge (\beta - \eta)\mathbf{1}.$$

For every action, $0 \le \bar{\phi}_i^{\theta}(h;s,a_i) \le B$, so the optimal-response operator satisfies

$$T_i^{\theta}h - h \le \big(\beta + (1-\gamma)B\big)\mathbf{1}.$$

Lemma C.5 bounds the played policy's payoff below by $\beta - \eta$ and every response's payoff above by $\beta + (1-\gamma)B$, under either nature convention. Subtracting proves the claim. Only these Bellman inequalities, not constant-gain solvability, are used. □

F Certificates for bounded optimal-response span

This section certifies the span assumption used in the equilibrium and learning results. We first reduce structural questions to fixed points of the zero-reward response operator and establish a bound uniform over opponents. We then prove the chain and hitting-time certificates and give the reward-uniform characterization. Appendix F.4 collects the proof of Theorem 4.2, its extensions to other divergences, and the localized TV certificates. The final examples separate the structural conditions.

F.1 Direct chain certificates

Throughout this section, local ambiguity sets may be arbitrary nonempty compact subsets of $\Delta(\mathcal{S})$, rectangular across primitive state–joint-action rows. The TV balls in the learning problem are one special case. The structural arguments suppress the nominal center and write $\mathcal{P}_{i,s,\boldsymbol{a}}$ for the local set. In the TV-specific results below, every ball is taken over the full state simplex and its radius lies in $[0,1]$; the learning model uses the smaller range $[0,1/2]$.

F.1.1 Notation and analytic preliminaries

Fix a player i and stationary opponents θ. In a proof for one response problem, suppress i,θ, write a for the responding player's action, and write $\mathcal{Q}_{s,a}$ for the induced row set $\mathcal{Q}^{i,\theta}_{s,a_i} \triangleq \big\{\sum_{a_{-i}} \theta(a_{-i} \mid s)p_{a_{-i}} : p_{a_{-i}} \in \mathcal{P}_{i,s,(a_i,a_{-i})}\big\}$. This set is the continuous image of a finite product of nonempty compact primitive ambiguity sets, so it is nonempty and compact. Primitive rectangularity gives

$$\min_{q\in\mathcal{Q}_{s,a}} q^{\top}v = \sum_{a_{-i}} \theta(a_{-i} \mid s) \min_{p\in\mathcal{P}_{i,s,(a,a_{-i})}} p^{\top}v.$$

Indeed, the right side is a lower bound for every choice of the primitive rows, and choosing a minimizer separately for each realized opponent action attains it. These choices can be made independently for every (s, a).

Subtract the minimum primitive reward of player i from every one-step reward. This shifts each discounted value by a constant and each gain by the same reward constant, without changing any span. The resulting induced rewards satisfy $0 \le r(s,a) \le R$, where $R = R_i$. Define

$$(T_\gamma v)(s) = \max_a \{r(s,a) + \gamma \min_{q \in \mathcal{Q}_{s,a}} q^\top v\}, \qquad (Fv)(s) = \max_a \min_{q \in \mathcal{Q}_{s,a}} q^\top v, \qquad T = T_1.$$

The discount satisfies $0 < \gamma < 1$, and $\lambda = 1 - \gamma > 0$. The standard discounted robust dynamic programming result, recalled in Proposition C.4, gives a unique fixed point v_γ of T_γ; no communication assumption is needed. The reward bounds give

$$0 \le v_\gamma \le (R/\lambda)\mathbf{1}. \tag{36}$$

The operator F is monotone, preserves constants, and satisfies $F(v + c\mathbf{1}) = Fv + c\mathbf{1}$ for $c \in \mathbb{R}$ and $F(bv) = bFv$ for $b \ge 0$. In particular, it is nonexpansive in span: $\mathrm{sp}(Fv - Fw) \le \mathrm{sp}(v - w)$. To see the last property, apply monotonicity and additive homogeneity to $w + \min(v - w)\mathbf{1} \le v \le w + \max(v - w)\mathbf{1}$. We will use these order properties only for the zero-reward operator; the reward-dependent operator T need not be positively homogeneous.

We use $H_i(\theta)$ and H_{R} from Definition C.1, and D_{R} from (23). All opponents range over the compact product of their statewise policy simplices. The primitive uncertainty sets are fixed when opponents vary.

Lemma F.1. *For a fixed response problem, suppose* $\sup_{0<\gamma<1} \mathrm{sp}(v_\gamma) \le K < \infty$. *Then* $\rho\mathbf{1} + h = Th$ *has a solution with* $\mathrm{sp}(h) \le K$. *Consequently, a bound* K *uniform over all response problems implies* $H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\{1, K\}$.

Proof. Apply the pointwise bounded-span implication already established in the proof of Theorem C.9, following (Wang & Si, 2025, Theorem 4), to the induced compact state–action rectangular response model. Its normalized discounted biases lie in $[0, K]^{\mathcal{S}}$, so the resulting bias has span at most K. The definitions give the uniform assertion. □

Lemma F.2. *Suppose* F_i^θ *has only constant fixed points for every* i, θ. *For* $S \ge 2$, *let*

$$\mathcal{S}_{\mathrm{sp}} = \{x : \min x = 0,\ \max x = 1\},$$
$$\eta_i(\theta) = \min_{x \in \mathcal{S}_{\mathrm{sp}}} \mathrm{sp}\big(x - F_i^\theta x\big), \qquad \eta_{\mathrm{R}} = \min_i \min_{\theta \in \Pi_{-i}} \eta_i(\theta).$$

Then $\eta_{\mathrm{R}} > 0$ *and*

$$D_{\mathrm{R}} \le \max\{1, 2R_{\mathrm{osc}}/\eta_{\mathrm{R}}\}. \tag{37}$$

In particular $H_{\mathrm{R}} < \infty$. *When* $S = 1$, $H_{\mathrm{R}} = D_{\mathrm{R}} = 1$ *directly; set* $\eta_{\mathrm{R}} = +\infty$ *in that case.*

Proof. The qualitative implication is the recession-operator criterion of (Akian et al., 2015, Theorem 2.2). We retain the quantitative compactness argument because the desired bound is uniform over opponents.

For normalized x, $0 \le F_i^\theta x \le 1$. If $x - F_i^\theta x = c\mathbf{1}$, evaluating at a minimum and a maximum of x gives $c \le 0$ and $c \ge 0$, respectively. Thus zero residual span would give a forbidden nonconstant fixed point. Moreover, $(\theta, x) \mapsto F_i^\theta x$ is continuous: primitive row minima are Lipschitz in x, and opponent weights, sums, and finite maxima are continuous. Compactness of $\Pi_{-i} \times \mathcal{S}_{\mathrm{sp}}$ and finiteness of the player set therefore give $\eta_{\mathrm{R}} > 0$.

We next turn this positive residual margin into a discounted-span bound. Fix a response problem and discount, and put $c = \min v_\gamma$, $h = v_\gamma - c\mathbf{1}$, and $B = \mathrm{sp}(v_\gamma)$. Because $0 \le r \le R_i$, the vector $e = T_\gamma h - \gamma Fh$ lies in $[0, R_i]^{\mathcal{S}}$. The normalized Bellman equation is $h + \lambda c\mathbf{1} = T_\gamma h$, hence $h - \gamma Fh = e - \lambda c\mathbf{1}$ and $\mathrm{sp}(h - \gamma Fh) \le R_i$. If $B > 0$, the vector $x = h/B$ belongs to $\mathcal{S}_{\mathrm{sp}}$ and

$$\eta_{\mathrm{R}} \le \mathrm{sp}(x - Fx) \le \frac{\mathrm{sp}(h - \gamma Fh) + \lambda\,\mathrm{sp}(Fh)}{B} \le \frac{R_i}{B} + \lambda.$$

The last inequality uses $\mathrm{sp}(Fh) \le B$. For $\lambda \le \eta_{\mathrm{R}}/2$, rearranging gives $B \le 2R_i/\eta_{\mathrm{R}}$. For $\lambda > \eta_{\mathrm{R}}/2$, (36) gives $B \le R_i/\lambda < 2R_i/\eta_{\mathrm{R}}$. The case $B = 0$ needs no division. Taking suprema proves (37), and Lemma F.1 gives existence. For $S = 1$ all spans vanish. □

Lemma F.3. *Let $\mathcal{R}_s$ be nonempty compact sets of available rows, and define $(Gx)(s) = \max_{q\in\mathcal{R}_s} q^\top x$. Suppose there is a nonempty set C such that every state outside C is transient under every stationary row selection, and every ordered pair of states in C is reachable under some stationary selection. Then $Gx = x$ implies that x is constant.*

Proof. Let $I = \arg\min x$ and $J = \arg\max x$. The equation $Gx = x$ makes I closed under every selection, while compactness provides one selection making J closed. Each nonempty closed set contains a recurrent class, hence both I and J meet C. Since I is closed under every decision and every state of C is reachable from a state of $I \cap C$, one has $C \subseteq I$. Thus $I \cap J \neq \varnothing$ and x is constant. This is the finite-state maximum principle underlying the weak-communication criterion. □

F.1.2 Proof of Theorem 4.1

Proof. The pointwise solvability assertions specialize (Wang & Si, 2025, Corollaries 7.1–7.2 and Theorems 6–7). We verify the zero-reward implication directly, so that the uniformity over opponents follows from Lemma F.2 without any additional uniform communication hypothesis.

Fix i, θ and a fixed point $Fx = x$. Compactness and rectangularity allow simultaneous minimizing rows q^x and a maximizing deterministic response a_x, with

$$x(s) = \max_a (q^x_{s,a})^\top x = \min_{q\in\mathcal{Q}_{s,a_x(s)}} q^\top x. \tag{38}$$

We show that each chain condition forces x to be constant.

Under condition 1, the selected matrix Q^{a_x,q^x} is unichain and $x = Q^{a_x,q^x}x$. Both the minimum and maximum sets of a harmonic vector are closed under this chain. If x were nonconstant, these disjoint nonempty closed sets would contain two different recurrent classes, contradicting the unichain assumption.

Under condition 2, choose the irreducible witness $q = q(a_x)$. The second equality in (38) gives $Q^{a_x,q}x \geq x$. At a state maximizing x, this inequality forces every positive-probability successor to have the same maximum value. The maximum set is therefore closed, and irreducibility makes it the entire state space.

Under condition 3, choose the irreducible response $\mu = \mu(q^x)$. The first equality in (38) gives $Q^{\mu,q^x}x \leq x$. Now the minimum set is closed, and irreducibility again implies constancy.

Under condition 4, (38) says that x is a fixed point of the maximizing controlled operator with rows $q^x_{s,a}$. Lemma F.3 applies to that weakly communicating model.

Thus every response operator has only constant fixed points, even if different response problems satisfy different conditions. Lemmas F.2 and F.1 give the uniform span bound and Bellman solutions. The argument concerns transitions only, and therefore applies separately to every bounded reward specification. □

Proposition F.4. *For every player i and stationary opponents θ, suppose there is a weakly communicating finite controlled reference model with rows $\bar{q}_{s,a_i}$ such that*

$$\bar{q}_{s,a_i}(t) > 0 \quad\Longrightarrow\quad q_{s,a_i}(t) > 0 \quad \textit{for every } q_{s,a_i} \in \mathcal{Q}^{i,\theta}_{s,a_i}. \tag{39}$$

Then every fixed-opponent response problem admits a constant-gain Bellman solution and $H_{\mathrm{R}} < \infty$. This is a reference-model certificate; the support-free condition in Theorem 4.1, item 4, is an alternative sufficient condition.

Proof. Fix $Fx = x$ and let I, J be its minimum and maximum sets. At $s \in I$, every action has a minimizing row supported on I; (39) therefore makes I closed under every reference action. At $s \in J$, one maximizing action has every admissible row supported on J, so its reference row is supported on J. Thus J is closed under one reference policy.

Both sets must meet the reference weak-communication core C, because a nonempty closed set contains a recurrent class and states outside C are universally transient. Closure of I under every reference action and reachability within C give $C \subseteq I$. Hence $I \cap J \neq \varnothing$, proving constancy. Lemmas F.2 and F.1 give the uniform conclusion. Only the support implication is used; no common numerical transition lower bound is required. □

Corollary F.5. *Assume the nominal row belongs to each ambiguity set. If every deterministic stationary joint policy induces an irreducible nominal chain, Condition 2 of Theorem 4.1 holds for every stationary opponent profile. For TV ambiguity, the primitive inequality $P^0_{s,\boldsymbol{a}}(t) > \sigma_{i,s,\boldsymbol{a}}$ preserves that edge under every admissible row. If all positive primitive nominal edges satisfy this inequality and every nominal fixed-opponent response MDP is weakly communicating, Proposition F.4 applies.*

Proof. For fixed opponents and a deterministic response, choose one supported opponent joint action at every state. Its nominal joint-policy chain is irreducible, and every edge of that chain remains positive in the opponent-averaged nominal kernel. This kernel is an admissible irreducible witness, proving the first claim.

For TV rows, $p(t) \geq P^0(t) - \sigma > 0$ by the event characterization of TV distance. A positive induced nominal edge has a positive-weight primitive witness, which remains positive under every admissible primitive selection. Thus (39) holds and Proposition F.4 applies. □

F.1.3 Weak communication, Bellman-active rows, and recurrent-class overlap

The next proposition collects the two one-sided weak-communication conditions and their Bellman-active refinement. Its second condition is item 4 of Theorem 4.1. Items 1–2 specialize the one-sided weak-communication theorems of (Wang & Si, 2025); item 3 records the Bellman-active refinement.

Proposition F.6. *Each of the following conditions, imposed on every fixed-opponent response problem, implies $H_{\mathrm{R}} < \infty$:*

1. *For every deterministic response a, the nature-controlled model with available rows $\mathcal{Q}_{s,a(s)}$ is weakly communicating in the sense of Lemma F.3.*
2. *For every full admissible row selection q, the controlled model with own-action rows $q_{s,a}$ is weakly communicating.*
3. *The condition in item 2 holds for every full selection whose rows simultaneously minimize the expectation of some common nonconstant vector. These are the Bellman-active row selections.*

The first two conditions are alternative sufficient conditions; the third weakens the second.

Proof. These are response-model specializations of the one-sided conditions in (Wang & Si, 2025, Theorems 5–7), with item 3 restricting the rows that must be checked. For the uniform conclusion, suppose $Fx = x$.

Under item 1, choose a_x as in (38). Then $x(s) = \min_{q \in \mathcal{Q}_{s,a_x(s)}} q^\top x$, so $-x$ is a fixed point of the maximizing nature-controlled operator. Its assumed weak communication and Lemma F.3 force x to be constant.

Under item 2, use the proof of condition 4 above. For item 3, if x were nonconstant, the simultaneous minimizing selection q^x in (38) would be Bellman-active. The same weak-communication argument would again force constancy. Lemma F.2 supplies the opponent-uniform bound in all three cases. □

Proposition F.7. *Consider any finite-state controlled model with independently selectable statewise decisions. Let $\mathfrak{R}$ be the family of all nonempty closed irreducible recurrent classes of all its stationary selections. Suppose any two members of $\mathfrak{R}$ are connected by a finite sequence of members with consecutive nonempty intersections. Then this model is weakly communicating with core $C = \bigcup_{R \in \mathfrak{R}} R$.*

Consequently, either of the following opponent-uniform conditions implies $H_{\mathrm{R}} < \infty$: for every deterministic response, this overlap property holds across all nature row selections; or for every full nature row selection, it holds across all stationary response policies.

Proof. This is the recurrent-class argument behind (Wang & Si, 2025, Corollary 7.3); we give the short statewise-selection argument to avoid that corollary's convexity assumptions. The set C is nonempty, and its complement contains no state recurrent under any stationary selection, hence is universally transient. For $s, t \in C$, an overlap chain of witnessing recurrent classes supplies a

directed walk from s to t. Erase loops to obtain a simple path. At each of its distinct states choose a decision assigning positive probability to the next state. Independent statewise selection extends these consistent choices to one stationary selection, which reaches t from s with positive probability. Thus C is a communicating core. Applying Proposition F.6 gives both consequences. The class family must include every recurrent class of every specified stationary selection. □

F.2 Hitting-time, reachability, and contraction certificates

This subsection converts quantitative reachability assumptions into uniform discounted-span bounds and constant-gain Bellman solutions. We first prove a directional comparison and the all-target theorem, then give common-anchor, finite row-mass, and contraction certificates.

F.2.1 Uniform hitting-time certificates

For a nonempty target set C, let $\tau_C = \inf\{k \geq 0 : S_k \in C\}$, and write $\tau_t = \tau_{\{t\}}$. Write ν_i for a possibly history-dependent response and ξ_i for an admissible history-dependent strategy of player i's nature. The expectation $\mathbb{E}_s^{\nu_i,\theta,\xi_i}$ starts at s, with opponents fixed at θ. Let $\mathcal{F}_k$ be the history just before the actions at time k. The responding player can use its history and current state, but not the opponents' current random draw. Nature observes the realized joint action before selecting its admissible row. These are the information conventions of the manuscript.

Lemma F.8. *Fix a response problem with translated rewards in $[0, R]$, a discount, and states s, t. If the player can ensure $\mathbb{E}_s\tau_t \leq D$ against every admissible nature strategy, then $v_\gamma(t) - v_\gamma(s) \leq RD$. If nature can ensure $\mathbb{E}_s\tau_t \leq D$ against every response strategy, then $v_\gamma(s) - v_\gamma(t) \leq RD$.*

Proof. This is the robust discounted version of the classical hitting-time comparison; see (Bartlett & Tewari, 2009, Theorem 4 and Corollary 5) for the non-robust bias bound. We retain the two Bellman directions to verify the adversarial quantifiers. Write $v = v_\gamma$ and $\lambda = 1 - \gamma$. Simultaneous primitive minimizers at v give a stationary kernel against which every response satisfies

$$\mathbb{E}[r_k + \gamma v(S_{k+1}) \mid \mathcal{F}_k] \leq v(S_k). \tag{40}$$

A deterministic Bellman-maximizing response gives, against every nature strategy,

$$v(S_k) \leq \mathbb{E}[r_k + \gamma v(S_{k+1}) \mid \mathcal{F}_k]. \tag{41}$$

These inequalities follow by averaging the primitive Bellman scores under the stated information order.

Here r_k is the translated realized one-step reward, so $0 \leq r_k \leq R$. For either choice of strategies, define $M_n = \sum_{k=0}^{n-1} \gamma^k r_k + \gamma^n v(S_n)$. Inequality (40) makes (M_n) a supermartingale under the minimizing kernel, whereas (41) makes it a submartingale under the maximizing response. Let $\tau = \tau_t$. Optional stopping at the bounded time $\tau \wedge n$ gives the corresponding expectation inequality. The forcing hypothesis gives $\tau < \infty$ almost surely, and $0 \leq M_{\tau\wedge n} \leq R/\lambda$ permits dominated convergence as $n \to \infty$. The player's forcing strategy against the minimizing kernel yields

$$v(s) \geq \mathbb{E}\left[\sum_{k<\tau} \gamma^k r_k + \gamma^\tau v(t)\right] \geq \mathbb{E}[\gamma^\tau] v(t).$$

Since $1 - \gamma^\tau \leq \lambda\tau$ and $v(t) \leq R/\lambda$, this gives $v(t) - v(s) \leq R\mathbb{E}\tau$. Nature's forcing strategy against the maximizing response yields the reversed stopped inequality and

$$v(s) \leq \mathbb{E}\left[\sum_{k<\tau} \gamma^k r_k + \gamma^\tau v(t)\right] \leq R\mathbb{E}\tau + v(t).$$

Each forcing hypothesis bounds its respective expectation by D. □

Theorem F.9. *Either of the following conditions is sufficient. The displayed constant must be finite and independent of i, θ, s, t.*

1. ***Player-forced reachability.*** *For every i, θ and ordered pair s, t, there exists a response ν_i such that*

$$\mathbb{E}_s^{\nu_i,\theta,\xi_i}[\tau_t] \leq D_{\mathrm{P}} \qquad \text{for every admissible } \xi_i.$$

2. ***Nature-forced reachability.*** *For every i, θ and ordered pair s, t, there exists an admissible strategy ξ_i such that*

$$\mathbb{E}_s^{\nu_i,\theta,\xi_i}[\tau_t] \le D_{\mathrm{N}} \qquad \text{for every response } \nu_i.$$

Set $D = D_{\mathrm{P}}$ under the first condition and $D = D_{\mathrm{N}}$ under the second. Then every response problem admits a constant-gain solution with $\mathrm{sp}\big(h_i^\theta\big) \le R_i D$, and

$$\mathrm{sp}\Big(V_{i,\gamma}^{\star,\theta}\Big) \le R_i D \quad (0 < \gamma < 1), \qquad H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\{1, R_{\mathrm{osc}} D\}. \tag{42}$$

Proof of Theorem F.9. Translate rewards into $[0, R_i]$. Apply Lemma F.8 from a minimum to a maximum of v_γ under player-forced reachability, and from a maximum to a minimum under nature-forced reachability. In either case $\mathrm{sp}(v_\gamma) \le R_i D$ for every discount. Lemma F.1 supplies a bias with the same span bound. Restoring the reward constant changes only the gain, and taking suprema over i, θ proves (42). □

The forcing strategy may depend on i, θ, s, t, but must work against every strategy of the other side. Only one of the two all-target conditions is needed. The proof uses discounted values before constructing an average-reward solution.

Divergence-ball applications. Appendix F.4 collects the TV and reverse-KL theorem proof and the other-divergence extensions. Their common target-probability condition also gives an all-target hitting-time certificate.

Corollary F.10. *For every i, θ, suppose there is a state $z_{i,\theta}$ such that, from every initial state s, both of the following hold:*

1. *some response ν_i ensures $\mathbb{E}_s^{\nu_i,\theta,\xi_i}\tau_{z_{i,\theta}} \le D_{\mathrm{P},z}$ for every admissible ξ_i;*
2. *some admissible ξ_i ensures $\mathbb{E}_s^{\nu_i,\theta,\xi_i}\tau_{z_{i,\theta}} \le D_{\mathrm{N},z}$ for every response ν_i.*

The constants $D_{\mathrm{P},z}, D_{\mathrm{N},z}$ are finite and uniform over i, θ, s. The anchor is common to both comparisons for the fixed response problem, but may depend on i, θ. Then every response problem admits a constant-gain solution with bias span at most $R_i(D_{\mathrm{P},z} + D_{\mathrm{N},z})$, and

$$H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\{1, R_{\mathrm{osc}}(D_{\mathrm{P},z} + D_{\mathrm{N},z})\}. \tag{43}$$

Proof. For translated rewards, the two parts of Lemma F.8 give

$$v_\gamma(z) - R_i D_{\mathrm{P},z} \le v_\gamma(s) \le v_\gamma(z) + R_i D_{\mathrm{N},z} \qquad (z = z_{i,\theta}).$$

This interval has length $R_i(D_{\mathrm{P},z} + D_{\mathrm{N},z})$. Take spans and apply Lemma F.1, then take suprema over response problems. The two strategies are used in separate comparisons and need not succeed simultaneously. □

F.2.2 Layered row-mass certificates

Fix i, θ. For a target set $K \subseteq \mathcal{S}$, define

$$b_{\mathrm{P}}(s, K) = \max_{a_i} \min_{q \in \mathcal{Q}_{s,a_i}^{i,\theta}} q(K), \qquad b_{\mathrm{N}}(s, K) = \min_{a_i} \max_{q \in \mathcal{Q}_{s,a_i}^{i,\theta}} q(K).$$

The first quantity is the largest one-step entry probability the player can guarantee against every nature row. The second is the probability nature can guarantee regardless of the own action, by choosing its primitive row after observing the realized joint action. The set $\mathcal{Q}_{s,a_i}^{i,\theta}$ averages these primitive choices over the opponents' draw; it does not give the responding player advance knowledge of that draw. A family of predecessor layers turns a positive row-mass margin into a finite expected hitting-time bound. This is a sufficient test; transitions need not decrease the layer index at every step.

Proposition F.11. *Fix a nonempty target set C and a response problem. For either $b = b_{\mathrm{P}}$ or $b = b_{\mathrm{N}}$, set*

$$K_0 = C, \qquad K_{k+1} = K_k \cup \{s \notin K_k : b(s, K_k) > 0\}.$$

Suppose $K_L = \mathcal{S}$ for some $L \geq 1$, and let

$$\beta = \min_{1 \leq k \leq L} \min_{s \in K_k \setminus K_{k-1}} b(s, K_{k-1}) > 0,$$

where empty layers are omitted. The corresponding side has a stationary strategy such that, against every history-dependent strategy of the other side,

$$\sup_s \mathbb{E}_s \tau_C \leq L/\beta^L. \tag{44}$$

If $C = \mathcal{S}$, the bound is zero without using L or β.

Proof. Give each state its first-entry layer as rank. At rank $k > 0$, in the player case choose an action maximizing $b_{\mathrm{P}}(s, K_{k-1})$; in the nature case choose, for each action, a row maximizing its mass on K_{k-1}. Compactness attains these maxima. In the latter case represent the induced row by its primitive rows and implement the corresponding primitive choice after the opponent action is observed; rectangularity makes all these choices compatible.

For either side, conditional on every pre-action history before the first hit, rank decreases with probability at least β, regardless of the other side's strategy. In the player case this is the opponent-averaged probability required by the information structure. Starting at any positive rank, at most L consecutive rank decreases reach the target. Multiplying these conditional lower bounds gives a hit within the next L steps with probability at least β^L. This argument does not require independence or monotone rank on unsuccessful steps. Restarting the bound after each unsuccessful block gives

$$\mathbb{P}_s(\tau_C > kL) \leq (1 - \beta^L)^k, \qquad \mathbb{E}_s \tau_C \leq L \sum_{k \geq 0} (1 - \beta^L)^k = L/\beta^L.$$

If $C = \mathcal{S}$, $\tau_C = 0$. □

Proposition F.12. *The following quantities are lower bounds for the corresponding fixed-opponent quantities, for every stationary opponent profile:*

$$\underline{b}_{\mathrm{P}}(s, K) = \max_{a_i} \min_{a_{-i}} \min_{p \in \mathcal{P}_{i,s,(a_i,a_{-i})}} p(K), \qquad \underline{b}_{\mathrm{N}}(s, K) = \min_{a_i, a_{-i}} \max_{p \in \mathcal{P}_{i,s,(a_i,a_{-i})}} p(K).$$

Thus Proposition F.11 applied to these quantities supplies opponent-uniform bounds. For a full-simplex TV ball and a nonempty proper set K,

$$\min_p p(K) = (P^0(K) - \sigma)_+, \qquad \max_p p(K) = \min\{1, P^0(K) + \sigma\}. \tag{45}$$

For $K = \varnothing$, both extrema are zero; for $K = \mathcal{S}$, both are one.

Proof. Rectangularity gives

$$b_{\mathrm{P}}(s, K) = \max_{a_i} \sum_{a_{-i}} \theta(a_{-i} \mid s) \min_p p(K), \quad b_{\mathrm{N}}(s, K) = \min_{a_i} \sum_{a_{-i}} \theta(a_{-i} \mid s) \max_p p(K).$$

Bounding each average below by its smallest summand proves the opponent-uniform inequalities. For (45), the TV event inequality bounds the change of $p(K)$ by σ. The endpoints are attained by transferring $\min\{\sigma, P^0(K)\}$ mass from K to K^c, or $\min\{\sigma, 1 - P^0(K)\}$ in the reverse direction. Both sets are nonempty, so these transfers are feasible in the full simplex. Boundary sets have fixed mass zero or one. □

Corollary F.13. *For every player, suppose there is a strongly connected directed graph on $\mathcal{S}$ and a number $\beta > 0$ such that each graph edge $s \to t$ has an own action a_i satisfying*

$$\min_{a_{-i}} \min_{p \in \mathcal{P}_{i,s,(a_i,a_{-i})}} p(t) \geq \beta.$$

For $S \geq 2$, this implies

$$H_{\mathrm{R}}, D_{\mathrm{R}} \leq \max\{1, R_{\mathrm{osc}}(S-1)/\beta^{S-1}\}.$$

Proof. Fix a target t. Rank states by their shortest directed path length to t. Strong connectivity makes each rank finite and at most $S - 1$. At every positive-rank state, choose the own action associated with the first edge of a shortest path. The process decreases rank with probability at least β, regardless of opponents and nature. Proposition F.11 supplies the player-forced all-target bound $(S-1)/\beta^{S-1}$. Theorem F.9 proves the result. The singleton case has zero hitting times and span parameters equal to one. □

F.2.3 A COMMUNICATING CORE WITH TRANSIENT ACCESS

Proposition F.14. *Fix a response problem and a nonempty set C. Suppose that every chain obtained from a deterministic stationary response and a full stationary admissible row selection reaches C in expected time at most T, uniformly over its initial state. Suppose, in addition, that one side can force every target $t \in C$ from every initial state in expected time at most D, against every strategy of the other side. Either side may have this forcing property. Then*

$$\sup_{0<\gamma<1} \operatorname{sp}(v_\gamma) \le R(D+T).$$

Uniform D, T over all response problems imply $H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\{1, R_{\mathrm{osc}}(D+T)\}$. Closure of C under all admissible rows is not required.

Proof. Fix a discount and take simultaneous Bellman-maximizing actions and minimizing rows at $v = v_\gamma$. The induced chain satisfies $v = r + \gamma Q v$, with $0 \le r \le R$. By the hypothesis its entrance time to C has mean at most T. The stopped identity from Lemma F.8, together with $1 - \gamma^{\tau_C} \le (1-\gamma)\tau_C$, gives

$$m_C - RT \le v(s) \le M_C + RT, \qquad m_C = \min_C v, \quad M_C = \max_C v. \tag{46}$$

Indeed, the stopped identity gives $v(s) \ge m_C \mathbb{E}_s[\gamma^{\tau_C}] \ge m_C - (1-\gamma) m_C \mathbb{E}_s \tau_C \ge m_C - RT$ and $v(s) \le R\mathbb{E}_s\tau_C + M_C \le M_C + RT$. These use $0 \le m_C \le M_C \le R/(1-\gamma)$. If the player forces targets in C, force a maximizer of v on C to get $v(s) \ge M_C - RD$ and combine with the upper bound in (46). If nature forces targets, force a minimizer to get $v(s) \le m_C + RD$ and combine with the lower bound. Both yield an interval of length $R(D+T)$. Lemma F.1 gives the uniform conclusion. □

Proposition F.15. *Let $\mathcal{R}_s$ be nonempty compact sets of rows in a finite state space, independently selectable across states, and let $C \neq \varnothing$. The following are equivalent:*

1. *Every stationary selection $q_s \in \mathcal{R}_s$ has every closed recurrent class intersecting C.*
2. *Starting from $K_0 = C$, the iteration*

$$K_{k+1} = K_k \cup \{s \notin K_k : \min_{q \in \mathcal{R}_s} q(K_k) > 0\}$$

reaches $\mathcal{S}$.

When these hold, all history-dependent choices from $\mathcal{R}_s$ reach C with a uniform finite expected hitting time. If L layers suffice and their smallest positive row-mass bound is β, the bound is L/β^L; if $C = \mathcal{S}$, it is zero.

Proof. If the iteration stops at a proper set K, put $D = \mathcal{S} \setminus K$. For each $s \in D$, compactness attains a row with $q_s(K) = 0$, so $q_s(D) = 1$. Choose these rows on D and arbitrary admissible rows elsewhere. The finite set D is closed and contains a recurrent class disjoint from C, contradicting the first condition.

Conversely, let a selected chain have a closed recurrent class D disjoint from C. Inductively suppose $D \cap K_k = \varnothing$. Its selected row at every $s \in D$ has mass zero on K_k, so the displayed minimum is zero and s is not added. This proves $D \cap K_k = \varnothing$ at every layer and prevents completion.

Under completion, choose the rank of first entry of each state. Every admissible row, and hence every history-dependent or randomized choice, decreases positive rank with probability at least β. The block argument of Proposition F.11 proves the claimed expectation bound. Randomizing among the rows does not change the minimum mass bound. □

For Proposition F.14, one may take $\mathcal{R}_s = \bigcup_a \mathcal{Q}_{s,a}$ in Proposition F.15; this is compact because the own action set is finite. For nominal models, take $\mathcal{R}_s$ to be the finite collection of primitive nominal joint-action rows. This yields a finite arithmetic check that upgrades a stationary recurrent-class condition to a bound valid against arbitrary history-dependent joint actions.

F.2.4 ADDITIONAL CONTRACTION CERTIFICATES

Assumption F.16 (Uniform robust minorization)**.** *There are $\alpha \in (0,1]$ and $\nu \in \Delta(\mathcal{S})$ such that every induced admissible response row q satisfies*

$$q \geq \alpha\nu$$

coordinatewise, for every player, opponent profile, state, own action, and admissible nature selection.

Proposition F.17. *Under Assumption F.16,*

$$\operatorname{sp}\big(F_i^\theta x\big) \leq (1-\alpha)\operatorname{sp}(x)$$

for all i, θ, x, and

$$D_{\mathrm{R}} \leq \max\left\{1, \frac{R_{\mathrm{osc}}}{\alpha}\right\}, \qquad H_{\mathrm{R}} \leq \max\left\{1, \frac{R_{\mathrm{osc}}}{\alpha}\right\}. \tag{47}$$

Proof. The common minorization is the classical Doeblin–Dobrushin span certificate; see (Gaubert & Qu, 2013, Theorem 1.1). Here its applicability to the optimized rows follows directly from

$$\alpha\nu^\top x + (1-\alpha)\min x \leq q^\top x \leq \alpha\nu^\top x + (1-\alpha)\max x.$$

Row minima and action maxima remain in this same interval. Thus $\operatorname{sp}\big(F_i^\theta x\big) \leq (1-\alpha)\operatorname{sp}(x)$. At the discounted fixed point,

$$\operatorname{sp}(v_\gamma) \leq R_i + \gamma(1-\alpha)\operatorname{sp}(v_\gamma), \qquad \operatorname{sp}(v_\gamma) \leq \frac{R_i}{1-\gamma(1-\alpha)} \leq \frac{R_i}{\alpha}.$$

Take suprema and apply Lemma F.1. The argument also covers $\alpha = 1$. □

One-step minorization can be replaced by contraction after several Bellman steps.

Proposition F.18. *Suppose there are an integer $m \geq 1$ and $\alpha \in (0,1]$ such that*

$$\operatorname{sp}\big((F_i^\theta)^m x - (F_i^\theta)^m y\big) \leq (1-\alpha)\operatorname{sp}(x-y) \tag{48}$$

for all i, θ, x, y. Then every F_i^θ has only constant fixed points and

$$\eta_{\mathrm{R}} \geq \frac{\alpha}{m}, \qquad D_{\mathrm{R}}, H_{\mathrm{R}} \leq \max\left\{1, \frac{mR_{\mathrm{osc}}}{\alpha}\right\}.$$

Proof. Write $F = F_i^\theta$. Nonexpansiveness and (48) give

$$\alpha\operatorname{sp}(x) \leq \operatorname{sp}(x - F^m x) \leq \sum_{k=0}^{m-1} \operatorname{sp}\big(F^k x - F^k(Fx)\big) \leq m\operatorname{sp}(x - Fx).$$

Hence every fixed point is constant and $\eta_{\mathrm{R}} \geq \alpha/m$. For the sharper discounted estimate, translate rewards to $[0, R_i]$ and write $T = T_{i,\gamma}^\theta$. Iterating $\gamma Fv \leq Tv \leq \gamma Fv + R_i\mathbf{1}$ using monotonicity and homogeneity gives

$$\gamma^m F^m v \leq T^m v \leq \gamma^m F^m v + R_i \sum_{j=0}^{m-1} \gamma^j \mathbf{1}.$$

At $v = v_\gamma = T^m v_\gamma$, use (48) with $y = 0$ to conclude

$$\operatorname{sp}(v_\gamma) \leq \frac{R_i \sum_{j=0}^{m-1} \gamma^j}{1-\gamma^m(1-\alpha)} \leq \frac{mR_i}{\alpha}.$$

Taking suprema and applying Lemma F.1 proves the stated bounds. □

F.3 Reward-uniform characterization and quantitative bias bounds

F.3.1 Closed regions preserved by the two sides

The direct certificates above can be related to an exact set-theoretic condition. The qualitative equivalence is a robust-response specialization of (Akian et al., 2015, Theorems 3.1 and 4.4). We verify that specialization and retain the quantitative margin and deterministic-opponent arguments.

Definition F.19. *Fix a response problem. A nonempty set I is a lower trap if*

$$\forall s \in I \; \forall a \; \exists q \in \mathcal{Q}_{s,a} : \quad q(I) = 1.$$

A nonempty set J is an upper trap if

$$\forall s \in J \; \exists a \; \forall q \in \mathcal{Q}_{s,a} : \quad q(J) = 1.$$

A lower trap can be kept closed by nature after each action. An upper trap can be kept closed by the responding player against every admissible row.

Theorem F.20. *For one fixed response problem, the following are equivalent:*

1. *F has only constant fixed points.*
2. *Every lower trap intersects every upper trap.*
3. *For every state reward perturbation $g \in \mathbb{R}^{\mathcal{S}}$, the equation $\rho\mathbf{1} + h = Th + g$ has a solution.*

If these equivalent properties hold for all players and stationary opponents, then $\eta_{\mathrm{R}} > 0$ and

$$H_{\mathrm{R}} \le \max\{1, R_{\mathrm{osc}}/\eta_{\mathrm{R}}\}, \qquad D_{\mathrm{R}} \le \max\{1, 2R_{\mathrm{osc}}/\eta_{\mathrm{R}}\}. \tag{49}$$

For this qualitative characterization, it suffices to check deterministic stationary opponent profiles. This reduction does not assert that the numerical minimum defining η_{R} is attained at such a profile.

Proof. For $S = 1$ all spans vanish and the assertions are immediate; assume $S \ge 2$.

Qualitative equivalence. Rectangularity and attainment give

$$(Th)(s) = \min_{q_s \in \prod_a \mathcal{Q}_{s,a}} \max_a \{r(s,a) + q_{s,a}^{\top} h\}.$$

This is a finite-state Shapley operator in the convention of (Akian et al., 2015): the nature action space is compact, the responding action space is finite, transitions are continuous, and rewards are bounded. Its recession operator is F, since $T(bx)/b \to Fx$. Their Theorem 3.1 gives $1 \Leftrightarrow 3$. Their Theorem 4.4 gives $1 \Leftrightarrow 2$, because the two invariant-set conditions in that theorem are exactly

$$F\mathbf{1}_{\mathcal{S}\setminus I} \le \mathbf{1}_{\mathcal{S}\setminus I}, \qquad F\mathbf{1}_J \ge \mathbf{1}_J.$$

Indeed, at $s \in I$ the first condition means every action has an attained row minimum of zero outside I, and at $s \in J$ the second means some action has every row supported on J. These are Definition F.19.

Quantitative bounds. Lemma F.2 gives $\eta_{\mathrm{R}} > 0$, discounted spans, and Bellman existence. Translate rewards into $[0, R_i]$. For any solution, $0 \le Th - Fh \le R_i\mathbf{1}$, so

$$\mathrm{sp}(h - Fh) = \mathrm{sp}(Th - Fh) \le R_i.$$

If $B = \mathrm{sp}(h) > 0$, normalizing $x = (h - \min h\mathbf{1})/B$ gives

$$B\eta_{\mathrm{R}} \le B\,\mathrm{sp}(x - Fx) = \mathrm{sp}(h - Fh) \le R_i.$$

The zero-span case is immediate. Taking suprema proves (49).

Deterministic opponents. An induced row is supported on I exactly when every positively weighted primitive row is supported on I. Rectangularity makes their existential choices independent. Thus the lower-trap condition requires, for every own action and every supported opponent joint action, some primitive row supported on I. The upper-trap condition requires one own action whose every primitive row is supported on J for every supported opponent joint action. At each state choose one opponent joint action with positive probability under θ, and fix each opponent to its corresponding component. Both requirements survive this restriction, so disjoint traps for mixed opponents also occur for deterministic opponents. The converse is immediate. This establishes qualitative certification only, without an extremality claim for η_{R}. □

The distinction between reward-specific and reward-uniform statements is essential. A disconnected transition structure can have $H_{\mathrm{R}} = 1$ for constant rewards even when its payment-free operator has nonconstant fixed points. Theorem F.20 characterizes structures that work for every bounded state-reward perturbation; the learning theorem only requires the actual reward-specific H_{R} to be finite.

F.4 Proof of Theorem 4.2 and divergence extensions

This subsection collects the TV and reverse-KL theorem proof, the extensions to other divergences, and the localized TV certificates. We invoke the preceding span and communication results and the cited divergence identities, recording their adaptations to game responses. The ambiguity sets remain rectangular across primitive rows, and nature observes the realized joint action. Write $q = P^0_{s,\boldsymbol{a}}$ for a primitive nominal row and $\kappa_{i,s,\boldsymbol{a}} \geq 0$ for its divergence radius. All radii are finite. Forward KL means $D_{\mathrm{KL}}(p\|q)$, and reverse KL means $D_{\mathrm{KL}}(q\|p)$, where $D_{\mathrm{KL}}(p\|q) = \sum_t p(t)\log(p(t)/q(t))$ with the usual extended-value conventions. A full-simplex ball has no additional restriction to the nominal support.

F.4.1 A direct bound from feasible target probabilities

Proposition F.21. *For each player, define*

$$\beta_i := \min_{s,\boldsymbol{a},t} \max_{p\in\mathcal{P}_{i,s,\boldsymbol{a}}} p(t).$$

If $\beta_i > 0$ for every player, then every fixed-opponent response problem has a scalar-gain Bellman solution with $\mathrm{sp}\big(h_i^\theta\big) \leq R_i/\beta_i$. Moreover,

$$\mathrm{sp}\Big(V_{i,\gamma}^{\star,\theta}\Big) \leq \frac{R_i}{1-\gamma+\gamma\beta_i}, \qquad H_{\mathrm{R}}, D_{\mathrm{R}} \leq \max\{1, \max_i R_i/\beta_i\}. \tag{50}$$

Proof. For any continuation vector v, compactness and the definition of β_i give, for each primitive row,

$$\inf_{p\in\mathcal{P}_{i,s,\boldsymbol{a}}} p^\top v \leq (1-\beta_i)\max v + \beta_i \min v.$$

This is the rowwise estimate used in Shi et al. (2023, Lemma 7 and Appendix C.3.1). It is preserved by averaging over opponents and maximizing over the responding player's actions. Their discounted-span argument therefore applies with coefficient β_i and reward range R_i, yielding the discounted bound in (50). Lemma F.1 supplies the scalar-gain solution and the uniform bounds on H_{R} and D_{R}. □

The row choices may depend on the target and realized joint action, so the bound is uniform over stationary opponents. Constant-gain solvability is a conclusion of the proposition.

F.4.2 Proof of the TV and reverse-KL theorem

Proof of Theorem 4.2. For any target t, consider $p_\alpha = (1-\alpha)q + \alpha\delta_t$. Then $p_\alpha(t) \geq \alpha$.

(i) For TV, $\mathrm{TV}(p_\alpha, q) = \alpha\,\mathrm{TV}(\delta_t, q) \leq \alpha$. Choosing $\alpha = \underline{\sigma}_i$ makes this row feasible for all primitive rows of player i.

(ii) For reverse KL, $p_\alpha \geq (1-\alpha)q$ implies $D_{\mathrm{KL}}(q\|p_\alpha) \leq -\log(1-\alpha)$. Choosing $\alpha = 1 - e^{-\underline{\kappa}_i}$ again gives a feasible row for every state and joint action.

Proposition F.21 with these player-specific coefficients proves both cases. The full-simplex assumption permits every target t, including states outside the nominal support. □

Qualitative constant-gain solvability also follows from the one-sided communication criterion of Wang & Si (2025, Corollary 7.1 and the following remark): replacing δ_t by a full-support distribution in the same mixtures produces admissible irreducible transitions. The preceding argument adds explicit span constants uniform over opponents.

F.4.3 Divergences that permit transitions outside nominal support

Besides reverse KL, consider squared Hellinger and reverse Pearson:

$$\mathsf{H}^2(p,q) := \sum_t (\sqrt{p(t)} - \sqrt{q(t)})^2, \qquad \chi^2(q\|p) := \sum_t \frac{(q(t)-p(t))^2}{p(t)}.$$

The Hellinger convention takes values in $[0,2]$. For reverse Pearson, a positive numerator over zero is infinite and $0/0 = 0$.

Corollary F.22. *Suppose all primitive sets are full-simplex balls of one of these three divergences. Put $\underline{\kappa}_i := \min_{s,\boldsymbol{a}} \kappa_{i,s,\boldsymbol{a}} > 0$ and $b_i := \min\{\underline{\kappa}_i, 2\}$. In Proposition F.21, valid lower bounds on β_i are*

divergence constraint	*lower bound on β_i*
$D_{\mathrm{KL}}(q\|p) \le \kappa_{i,s,\boldsymbol{a}}$	$1 - e^{-\underline{\kappa}_i}$
$\mathsf{H}^2(p,q) \le \kappa_{i,s,\boldsymbol{a}}$	$b_i - b_i^2/4$
$\chi^2(q\|p) \le \kappa_{i,s,\boldsymbol{a}}$	$\underline{\kappa}_i/(1+\underline{\kappa}_i)$.

No nominal communication or full-support condition is required.

Proof. The reverse-KL row is case (ii) of Theorem 4.2. For the remaining rows use $p_\alpha = (1-\alpha)q + \alpha\delta_t$. The bounds

$$\sum_j \sqrt{q(j)p_\alpha(j)} \ge \sqrt{1-\alpha}, \qquad \chi^2(q\|p_\alpha) = \sum_{j:p_\alpha(j)>0} \frac{q(j)^2}{p_\alpha(j)} - 1 \le \frac{\alpha}{1-\alpha}$$

give the displayed Hellinger and reverse-Pearson coefficients. At Hellinger radius at least two, δ_t itself is feasible. Apply Proposition F.21. □

Example F.23 (Sharpness of the span constants)**.** *Take two states $0, 1$, one action, nominal identity transitions and rewards $r(0) = 0$, $r(1) = R > 0$. Use a common positive radius for either reverse KL, Hellinger or reverse Pearson. Let α be its value in Corollary F.22. Nature can keep state 0 absorbing and move from 1 to 0 with probability α. This is the largest possible such probability, because the divergences from a Dirac nominal row are respectively $-\log(1-\alpha)$, $2(1-\sqrt{1-\alpha})$ and $\alpha/(1-\alpha)$. The robust gain is zero and the Bellman equation is solved by $h(0) = 0$, $h(1) = R/\alpha$. Every solution has this span: the equation at state 1 first forces $h(1) > h(0)$ and then gives $R = \alpha(h(1) - h(0))$. Thus the three universal span constants are attained exactly. Duplicating actions embeds the construction into a finite game.*

F.4.4 Forward KL and Pearson with full-support nominal rows

Corollary F.24. *If all primitive nominal rows have full support and belong to their ambiguity sets, define $q_{\min} := \min_{s,\boldsymbol{a},t} P^0_{s,\boldsymbol{a}}(t) > 0$. Then $H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\{1, R_{\mathrm{osc}}/q_{\min}\}$. This applies in particular to forward-KL and forward-Pearson balls of any finite nonnegative radii, where $\chi^2(p\|q) = \sum_t (p(t)-q(t))^2/q(t)$.*

Proof. The center is feasible, so $\beta_i \ge q_{\min}$. Apply Proposition F.21. □

The radii can be incorporated through the standard binary aggregation of f-divergences (Csiszár & Shields, 2004, Lemma 4.1 and Section 5.1, Eq. (5.1)). For $0 < a < 1$, set

$$d(z\|a) := z\log(z/a) + (1-z)\log((1-z)/(1-a)), \qquad u_{\mathrm{KL}}(a,\kappa) := \max\{z \in [a,1] : d(z\|a) \le \kappa\}.$$

For a forward-KL row, its exact maximum mass on target t is $u_{\mathrm{KL}}(q(t),\kappa)$. This equals one when $\kappa \ge \log(1/q(t))$, and otherwise is obtained by solving one monotone scalar equation. The endpoint values are $u_{\mathrm{KL}}(0,\kappa) = 0$ and $u_{\mathrm{KL}}(1,\kappa) = 1$. For forward Pearson, the corresponding exact quantity is

$$u_{\chi^2}(a,\kappa) = \min\{1, a + \sqrt{\kappa a(1-a)}\} \quad (0 < a < 1),$$

with endpoint values zero and one. Taking the minimum of these quantities over all primitive rows and targets gives β_i. These are the binary KL and Pearson costs from the cited aggregation identity; equality is attained by preserving the nominal conditional law on the complement of the target.

For a closed-form forward-KL bound, suppose $S \geq 2$ and put $\underline{\kappa} = \min_{i,s,\boldsymbol{a}} \kappa_{i,s,\boldsymbol{a}}$ and $\alpha = \min\{1, \underline{\kappa}/\log(1/q_{\min})\}$. Convexity yields $D_{\mathrm{KL}}((1-\alpha)q + \alpha\delta_t \| q) \leq \alpha \log(1/q(t))$. Hence $q_{\min} + (1-q_{\min})\alpha$ can replace $q_{\min}$ in Corollary F.24. More generally, if every deterministic nominal joint-policy chain is irreducible, Corollary F.5 proves $H_{\mathrm{R}} < \infty$ at all radii for any ambiguity family containing its nominal rows.

F.4.5 Preserving a weakly communicating response model

Proposition F.25. *Suppose every nominal fixed-opponent response MDP is weakly communicating. For a primitive row q, preserving every positive coordinate is sufficient for $H_{\mathrm{R}} < \infty$. The following strict rowwise radius bounds ensure this preservation:*

constraint	$\kappa_{i,s,\boldsymbol{a}} < \min_{t:q(t)>0} \tau(q(t))$
$D_{\mathrm{KL}}(p\|q) \leq \kappa_{i,s,\boldsymbol{a}}$	$\tau(a) = -\log(1-a)$
$\chi^2(p\|q) \leq \kappa_{i,s,\boldsymbol{a}}$	$\tau(a) = a/(1-a)$
$\mathsf{H}^2(p,q) \leq \kappa_{i,s,\boldsymbol{a}}$	$\tau(a) = 2(1-\sqrt{1-a})$.

At $a = 1$, the first two thresholds are $+\infty$ and the third is two. Reverse-KL balls preserve all positive nominal coordinates at every finite radius. This remains true if the balls are additionally restricted to the nominal support.

Proof. For $0 < a = q(t) < 1$, the smallest divergence subject to $p(t) = 0$ is the binary divergence with masses $(0,1)$ and $(a, 1-a)$, by the aggregation identity cited above. Substitution gives the three displayed thresholds, attained by conditioning q on t^c. The endpoint cases follow from the definitions. For reverse KL, $q(t) > 0$ and $p(t) = 0$ have infinite cost. Thus each strict bound preserves every positive primitive nominal edge. A positive induced nominal edge has at least one positive-weight primitive witness, whose probability stays positive under every admissible primitive selection. Proposition F.4 then proves constant-gain solvability and uniform span finiteness. □

Lemma F.2 and (49) give the corresponding residual-margin bounds, uniformly over the entire opponent policy simplices.

For a support-restricted lower semicontinuous convex f-divergence $D_f(p\|q) = \sum_{j:q(j)>0} q(j) f(p(j)/q(j))$ with $f(1) = 0$, the same argument gives the deletion cost

$$\tau_f(a) = af(0) + (1-a)f(1/(1-a)), \qquad 0 < a < 1.$$

Set $\tau_f(1) = +\infty$, since a support-restricted singleton row cannot change. This is again the binary aggregation formula (Csiszár & Shields, 2004, Lemma 4.1). Radii below all positive-coordinate deletion costs give the same communication certificate.

F.4.6 Exact TV and forward-KL closed-region tests

Corollary F.26. *For full-simplex TV balls, a nonempty set I is a lower trap exactly when*

$$P^0_{s,(a_i,a_{-i})}(I^c) \leq \sigma_{i,s,(a_i,a_{-i})}$$

for every $s \in I$, every own action, and every supported opponent action. A nonempty proper set J is an upper trap exactly when, at every $s \in J$, there is an own action such that every supported opponent action satisfies

$$\sigma_{i,s,(a_i,a_{-i})} = 0, \qquad P^0_{s,(a_i,a_{-i})}(J) = 1.$$

Proof. For a proper nonempty I, (45) gives $\max_p p(I) = 1$ exactly when $P^0(I^c) \leq \sigma$; $I = \mathcal{S}$ is automatically a lower trap. For proper nonempty J, $\min_p p(J) = 1$ exactly when $\sigma = 0$ and $P^0(J) = 1$. The full state space is always an upper trap, irrespective of the radii, which is why the upper-trap formula is stated only for proper J. Apply the primitive lower- and upper-trap conditions established in the deterministic-opponent part of Theorem F.20. □

Proposition F.27. *Fix i, θ. For forward-KL balls, a nonempty set I is a lower trap exactly when, for every $s \in I$, every a_i and every a_{-i} with $\theta(a_{-i} \mid s) > 0$,*

$$P^0_{s,(a_i,a_{-i})}(I) \geq e^{-\kappa_{i,s,(a_i,a_{-i})}}. \tag{51}$$

A nonempty set J is an upper trap exactly when, at every $s \in J$, there is an a_i such that for every supported a_{-i},

$$P^0_{s,(a_i,a_{-i})}(J) = 1. \tag{52}$$

Consequently, the condition that for every player and every deterministic stationary opponent profile there are no disjoint sets satisfying these tests is equivalent to reward-uniform scalar Bellman solvability and implies $H_{\mathrm{R}} < \infty$ for every bounded reward specification.

Proof. The cited aggregation identity gives minimum KL cost $-\log q(I)$ for a row supported on I, attained by $q(\cdot \mid I)$ when $q(I) > 0$. If $q(I) = 0$, no such row has finite KL cost. This proves (51). All feasible rows are supported on $\operatorname{supp} q$ and the center is feasible, so every feasible row is supported on J if and only if $q(J) = 1$. This proves (52). Apply Theorem F.20, including its reduction to deterministic opponents and its uniform compactness argument. □

The test is exact for solvability for every reward perturbation. For one prescribed reward function, a Bellman solution can exist even when the test fails. Thus its failure alone does not establish $H_{\mathrm{R}} = +\infty$ for the given rewards.

F.4.7 TV CONTRACTION AND BIAS UNIQUENESS

Lemma F.28. *Let $p \in \Delta(\mathcal{S})$, $0 \le \sigma \le 1$, and $m(v) = \min_s v(s)$. Then*

$$\phi_{p,\sigma}(v) = \sigma m(v) + \min_{u \in \mathcal{U}(p,\sigma)} u^\top v, \qquad \mathcal{U}(p,\sigma) = \{u : 0 \le u \le p,\ \mathbf{1}^\top u = 1 - \sigma\}. \tag{53}$$

Consequently, if every primitive TV radius in a response problem is at least $\alpha > 0$, then

$$\operatorname{sp}(T_\gamma v - T_\gamma w) \le \gamma(1-\alpha)\operatorname{sp}(v - w) \qquad (0 < \gamma \le 1). \tag{54}$$

The conclusion at $\gamma = 1$ concerns the undiscounted operator T.

Proof. Identity (53) is the primal form of the standard TV duality formula in Lemma G.4; see also Shi et al. (2024b, Appendix A.3, Lemma 1). We use it to obtain the contraction for heterogeneous primitive radii and fixed opponents.

For a primitive radius $\sigma_x \ge \alpha$, write

$$\phi_x(v) = \alpha m(v) + \psi_x(v), \qquad \psi_x(v) = (\sigma_x - \alpha)m(v) + \min_{u \in \mathcal{U}(p_x,\sigma_x)} u^\top v.$$

The map ψ_x is monotone and obeys $\psi_x(v + c\mathbf{1}) = \psi_x(v) + (1-\alpha)c$. Thus opponent averaging and own-action maximization give $T_\gamma v = \gamma\alpha m(v)\mathbf{1} + G_\gamma v$, where G_γ is monotone and has translation coefficient $\gamma(1-\alpha)$. Applying these properties to $w + \min(v-w)\mathbf{1} \le v \le w + \max(v-w)\mathbf{1}$ proves (54); the removed term is constant across states. □

Corollary F.29. *When all primitive radii of a fixed response problem have a positive common lower bound, its constant-gain Bellman solution has a unique bias modulo constants and a unique gain.*

Proof. Existence is established above. For two solutions, (54) at $\gamma = 1$ gives

$$\operatorname{sp}\Big(h - \widetilde{h}\Big) = \operatorname{sp}\Big(Th - T\widetilde{h}\Big) \le (1-\alpha)\operatorname{sp}\Big(h - \widetilde{h}\Big).$$

Hence their biases differ by a constant, and additive homogeneity in the two Bellman equations forces equality of the gains. □

F.4.8 LOCALIZED TV: ACTION-DEPENDENT RECURRENT CLASSES

The remaining certificates allow some primitive radii to vanish. Proposition F.30 treats positive-radius access in every nominal recurrent class; Proposition F.31 requires it only in proper classes and gives the target-initialized finite test. These statements retain the localized TV extension separately from Theorem 4.2.

Proposition F.30. *Fix a player i in a full-simplex TV game. The following conditions are equivalent:*

1. *For every deterministic stationary joint policy $\boldsymbol{a}$, each closed recurrent class of its nominal chain contains a state s with $\sigma_{i,s,\boldsymbol{a}(s)} > 0$.*
2. *The following iteration reaches $\mathcal{S}$:*

$$K_0 = \varnothing, \quad K_{k+1} = K_k \cup \left\{ s \notin K_k : \text{for every } \boldsymbol{a}, \quad \sigma_{i,s,\boldsymbol{a}} > 0 \text{ or } P^0_{s,\boldsymbol{a}}(K_k) > 0 \right\}. \tag{55}$$

If they hold, let $L_i \le S$ be the number of rounds required, and define

$$\sigma_{i,+} = \min\{\sigma_{i,s,\boldsymbol{a}} : \sigma_{i,s,\boldsymbol{a}} > 0\}, \qquad p_+ = \min\{P^0_{s,\boldsymbol{a}}(t) : P^0_{s,\boldsymbol{a}}(t) > 0\},$$
$$\beta_i = \min\{\sigma_{i,+}, p_+\} > 0.$$

For every target t, nature has a stationary admissible row selection that hits t in expected time at most $L_i/\beta_i^{L_i}$ against every joint-action strategy, including history-dependent strategies. If the condition holds for every player, then

$$H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\left\{1, \max_i R_i \frac{L_i}{\beta_i^{L_i}}\right\}. \tag{56}$$

Proof. If (55) stops at a proper K, each state in $D = \mathcal{S} \setminus K$ has an action with zero radius and nominal row supported on D. These statewise actions form a stationary joint policy with a recurrent class in D having only zero radii. Conversely, any such recurrent class is disjoint from every layer by induction. This proves the equivalence.

Under completion at least one radius is positive, because the first nonempty layer requires positive radii for all its actions; hence $\sigma_{i,+}$ and β_i are positive. For target t, use

$$q^{(t)}_{i,s,\boldsymbol{a}} = (1 - \sigma_{i,s,\boldsymbol{a}}) P^0_{s,\boldsymbol{a}} + \sigma_{i,s,\boldsymbol{a}} \delta_t. \tag{57}$$

Its TV distance from the center is at most $\sigma_{i,s,\boldsymbol{a}}$. Before hitting t, every realized action either has positive radius and hits t next with probability at least $\sigma_{i,+}$, or has zero radius and decreases the first-entry rank with nominal probability at least p_+. Therefore each step hits t or lowers rank with conditional probability at least β_i. After L_i consecutive such steps the target must be hit, since rank one has no lower nonempty layer. The block calculation of Proposition F.11 gives $\mathbb{E}\tau_t \le L_i/\beta_i^{L_i}$ against arbitrary joint-action histories. Theorem F.9 now yields (56). □

Proposition F.31. *Fix a player i and $S \ge 2$. The following three conditions are equivalent:*

1. *Every proper closed recurrent class of every deterministic stationary nominal joint-policy chain contains a state with positive selected radius for player i.*
2. *For every target t, the iteration*

$$K_0(t) = \{t\}, \quad K_{k+1}(t) = K_k(t) \cup \left\{ s \notin K_k(t) : \ \forall \boldsymbol{a}, \quad \sigma_{i,s,\boldsymbol{a}} > 0 \text{ or } P^0_{s,\boldsymbol{a}}(K_k(t)) > 0 \right\} \tag{58}$$

   reaches $\mathcal{S}$.
3. *For every target, nature can ensure a uniformly finite expected hitting time against all history-dependent joint-action strategies.*

Set $\sigma_{i,+}$ to the smallest positive primitive radius if one exists, and to one if all radii vanish. With p_+ as in Proposition F.30, put $\beta_i = \min\{\sigma_{i,+}, p_+\} > 0$. Under these conditions, if $L_{i,t} \le S - 1$ rounds suffice for target t, nature can use (57) *to obtain*

$$\sup_s \sup_{\text{joint-action strategies}} \mathbb{E}_s \tau_t \le \frac{L_{i,t}}{\beta_i^{L_{i,t}}} \le \frac{S-1}{\beta_i^{S-1}}.$$

If the conditions hold for every player, then

$$H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\left\{1, \max_i R_i \frac{S-1}{\beta_i^{S-1}}\right\}. \tag{59}$$

For $S = 1$, hitting times and spans are zero and $H_{\mathrm{R}} = D_{\mathrm{R}} = 1$ without any radius condition.

Proof. For a fixed target t, a proper terminal set in (58) leaves a closed complement whose states admit zero-radius nominal actions. A recurrent class in that complement is proper and violates condition 1. Conversely, a proper zero-radius recurrent class excludes some target t and, by induction, never enters its layers. This proves $1 \Leftrightarrow 2$.

Under condition 2 use (57). Every action either resets to the target with probability at least $\sigma_{i,+}$ or has zero radius and lowers rank with nominal probability at least p_+. Proposition F.11 gives $L_{i,t}/\beta_i^{L_{i,t}} \le (S-1)/\beta_i^{S-1}$, proving condition 3. If all radii vanish, only the second alternative is used.

If condition 1 fails, the witnessing stationary joint policy confines every nature strategy to its proper zero-radius recurrent class, so no external target can be reached. Thus $3 \Rightarrow 1$. Theorem F.9 proves (59); the singleton case is immediate. □

This condition permits a zero-radius recurrent class equal to the entire state space. For example, if all radii vanish and every deterministic nominal joint-policy chain is irreducible, it holds. Proposition F.30 gives a stronger, single-pass certificate starting from the empty set; Proposition F.31 uses one target-initialized pass per state to cover the additional irreducible zero-radius case.

F.4.9 LOCALIZED TV: AN UNAVOIDABLE RESET SET

Proposition F.32. *For player i, let $Z_i \neq \varnothing$ and $\alpha_i > 0$ satisfy $\sigma_{i,s,\boldsymbol{a}} \ge \alpha_i$ for every $s \in Z_i$ and every joint action. Suppose that under the nominal transitions, every history-dependent joint-action strategy reaches Z_i from every initial state in expected time at most $T_i < \infty$. Then nature can force every target in expected time at most $(T_i+1)/\alpha_i$, against every response and every stationary opponent profile. If such sets exist for all players, every fixed-opponent response problem admits a constant-gain solution and*

$$H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\left\{1, \max_i \frac{R_i(T_i+1)}{\alpha_i}\right\}. \tag{60}$$

A sufficient finite chain test for the existence of a uniform T_i is that every closed recurrent class of every deterministic stationary nominal joint-policy chain intersects Z_i.

Proof. Nature follows nominal rows until reaching Z_i, then uses $q = (1-\alpha_i)P^0_{s,\boldsymbol{a}} + \alpha_i \delta_t$, whose TV distance from its center is at most α_i. Represent the mixture by an independent reset coin of success probability α_i. After each failure, the conditional expected nominal travel time back to Z_i is at most T_i, and the reset transition adds one step. Consequently

$$\mathbb{E}\tau_t \le (T_i+1)\sum_{k\ge 1}(1-\alpha_i)^{k-1} = \frac{T_i+1}{\alpha_i}.$$

Incidental visits to t only shorten this bound. The hypothesis permits arbitrary continuation strategies after each history, so the estimate is uniform over joint-action strategies. Apply Theorem F.9. For the final assertion apply Proposition F.15 to the finite row sets $\{P^0_{s,\boldsymbol{a}} : \boldsymbol{a} \in \mathcal{A}\}$ and target Z_i. □

The two TV reset certificates have different scopes. Proposition F.30 permits the reset radius to depend on the chosen joint action, and checks precisely whether a zero-radius closed nominal selection exists. Proposition F.32 is often more convenient when physical states have reset authority for every action. Its T_i can be bounded directly, instead of using the potentially conservative minimum-probability estimate in (56).

Implications for this paper. These results certify the hypothesis used for equilibrium existence and the deterministic discounted-to-average reduction. Their proof combines the one-sided communication principle (Wang & Si, 2025) with explicit divergence geometry and uniform control over opponents. Related KL and f-divergence learning results use nominal uniform ergodicity and small uncertainty radii (Chen et al., 2025). Extending the TV statistical rate requires a separate analysis of empirical-center perturbations for the chosen divergence.

F.5 Examples separating the structural conditions

Example F.33 (Joint accessibility disappears after fixing opponents)**.** *Let $\mathcal{S} = \{0, 1\}$. Player 1 has one action. Player 2 has actions b_{stay} and b_{flip}. Under b_{stay}, both states self-loop. Under b_{flip}, the states alternate. The union joint-action graph is strongly connected, and a cooperative joint policy communicates.*

Give player 1 rewards $r_1(0) = 0$ and $r_1(1) = 1$. Fix player 2 to b_{stay}. Player 1 has no remaining transition control, and

$$V_\gamma(0) = \gamma V_\gamma(0), \qquad V_\gamma(1) = 1 + \gamma V_\gamma(1).$$

Thus

$$V_\gamma(0) = 0, \qquad V_\gamma(1) = \frac{1}{1-\gamma}, \qquad \mathrm{sp}(V_\gamma) = \frac{1}{1-\gamma} \to \infty.$$

The payment-free fixed-opponent response operator F is the identity (the reward-bearing operator is $Tv = r + v$). Communication paths that require the opponent to change its policy cannot be used by a unilateral deviator.

Example F.34 (Nominal unichain is not a sufficient certificate)**.** *Consider one action and two states $0, 1$, with rewards $r(0) = 1$, $r(1) = 0$. Fix $0 < \eta \le 1/2$ and let*

$$P_0^0 = \delta_0, \quad \sigma_0 = 0, \qquad P_1^0 = \eta\delta_0 + (1-\eta)\delta_1, \quad \sigma_1 = \eta.$$

The nominal chain has unique recurrent class $\{0\}$, and state 1 is transient. Nevertheless $H_{\mathrm{R}} = +\infty$.

Proof. The row δ_1 is admissible because $\mathrm{TV}(\delta_1, P_1^0) = \eta$. Thus $v(0) = 1/(1-\gamma)$, $v(1) = 0$ satisfies the discounted Bellman equation and, by contraction, is its value; the span diverges. A constant-gain solution is also impossible directly: state 0 forces $\rho = 1$, whereas the feasible row δ_1 at state 1 forces $\rho + h(1) \le h(1)$, hence $\rho \le 0$. Singleton opponent action sets embed the example in a game. □

Example F.35 (Positive TV regularizes an identity nominal kernel, sharply)**.** *Let every state have one action, $P_s^0 = \delta_s$, and the same full-simplex TV radius $\alpha \in (0, 1/2]$. For any reward vector r, write $r_{\min} = \min_s r(s)$ and $R = \max_s r(s) - r_{\min}$. Then*

$$\rho = r_{\min}, \qquad h(s) = \frac{r(s) - r_{\min}}{\alpha}$$

is the normalized constant-gain solution, and its span is R/α. The discounted span is

$$\mathrm{sp}(v_\gamma) = \frac{R}{1 - \gamma + \gamma\alpha}.$$

Thus the positive-radius bound is attained even though the nominal identity chain has S recurrent classes and remains admissible.

Proof. Lemma F.28 gives $\phi_s(v) = (1-\alpha)v(s) + \alpha \min v$. Substitution of the displayed h, whose minimum is zero, reduces the Bellman equation to $\rho + \alpha h(s) = r(s)$. Corollary F.29 gives uniqueness modulo constants. For the discounted equation, writing $m = \min v_\gamma$ gives

$$(1 - \gamma + \gamma\alpha)v_\gamma(s) = r(s) + \gamma\alpha m, \qquad (1-\gamma)m = r_{\min}.$$

Subtract the minimum equation to obtain the displayed span formula. If $R = 0$, all spans vanish and the defining floor gives $H_{\mathrm{R}} = 1$. □

Example F.36 (Disconnected nominal cycles with localized positive radii)**.** *Let the nominal transitions be the two deterministic cycles $1 \leftrightarrow 2$ and $3 \leftrightarrow 4$, with one action. Put radius $\alpha \in (0, 1/2]$ at states 2 and 4, and radius zero at states 1 and 3. Then for every reward range R,*

$$H_{\mathrm{R}}, D_{\mathrm{R}} \le \max\{1, 2R/\alpha\}.$$

Proof. Take the reset set $Z = \{2, 4\}$. Every nominal initial state reaches Z in at most one step, so $T = 1$. Every row at Z has radius α. Proposition F.32 gives the displayed bound. The nominal chain is multichain and contains no path between the two cycles; nature's admissible resets supply that connection. The zero radii on the other two states are compatible with the bound. □

Example F.37 (Two incompatible multichain decompositions)**.** *Let $\mathcal{S} = \{1, 2, 3, 4\}$, one action be available, and*

$$\mathcal{P}_1 = \mathrm{conv}\{\delta_2, \delta_4\}, \qquad \mathcal{P}_2 = \mathrm{conv}\{\delta_1, \delta_3\},$$
$$\mathcal{P}_3 = \mathrm{conv}\{\delta_2, \delta_4\}, \qquad \mathcal{P}_4 = \mathrm{conv}\{\delta_1, \delta_3\}.$$

The payment-free operator is

$$(Fx)(1) = (Fx)(3) = \min\{x(2), x(4)\},$$
$$(Fx)(2) = (Fx)(4) = \min\{x(1), x(3)\}.$$

If $Fx = x$, then $x(1) = x(3) = a$ and $x(2) = x(4) = b$. The fixed-point equations give $a = b$; hence every fixed point is constant.

The ambiguity set contains the kernel

$$P^A : 1 \leftrightarrow 2, \qquad 3 \leftrightarrow 4,$$

whose recurrent classes are $\{1, 2\}$ and $\{3, 4\}$, and the kernel

$$P^B : 1 \leftrightarrow 4, \qquad 2 \leftrightarrow 3,$$

whose recurrent classes are $\{1, 4\}$ and $\{2, 3\}$. Both kernels are multichain. The four displayed recurrent classes are overlap connected:

$$\{1, 2\} \leftrightarrow \{2, 3\} \leftrightarrow \{3, 4\} \leftrightarrow \{4, 1\}.$$

Every recurrent class of any admissible kernel contains an odd state and an even state, since every transition switches parity. It therefore intersects at least one of the four displayed classes. The entire recurrent-class collection is consequently overlap-connected, and Proposition F.7 applies as well. The direct fixed-point calculation above already establishes the conclusion without relying on this certificate.

Example F.38 (Player bridge actions with zero ambiguity)**.** *Consider a one-player model with $S \geq 2$ states on a directed cycle and primitive reward oscillation R. At each state the player can either remain there or move to the next state deterministically. There is no ambiguity. The model may also be embedded in a game by adding players with singleton action sets and constant rewards; their response spans are zero. Then*

$$H_{\mathrm{R}}, D_{\mathrm{R}} \leq \max\{1, R(S-1)\}.$$

The stationary policy that always remains has S recurrent classes.

Proof. For any ordered pair of states, choosing the bridge action until the target is reached takes at most $S - 1$ steps. This is a player-forced all-target bound with $D_{\mathrm{P}} = S - 1$. Theorem F.9 applies. The always-remain policy makes every state absorbing, so a bound for the optimal-response span does not require every response policy to be unichain. □

Example F.39 (Why the full-simplex scope matters)**.** *Positive numerical TV radii do not by themselves imply bounded span when every ambiguity row is restricted to its nominal successor face.*

Proof. Take the identity nominal kernel and restrict the row at state s to distributions supported on $\{s\}$. This face contains only δ_s, so every face-restricted TV ball is the same singleton, regardless of its positive radius. With unequal rewards at two states, their discounted value difference is the reward difference divided by $1 - \gamma$. Moreover a constant-gain equation would require $\rho = r(s)$ at every absorbing state, which is impossible for unequal rewards. Thus $H_{\mathrm{R}} = +\infty$. The reset and trimming arguments above require the full state simplex exactly where they transfer mass to a possibly nonnominal successor. □

G Upper bound for discounted plug-in learning

This section proves the learning guarantee in three parts. First, one coordinatewise Bernstein event controls every continuation vector, including vectors selected from the transition data. Second, an interpolation of nominal centers converts value perturbations into Bellman variance terms. Third,

a forward optimal-response path and a backward selected-response path give the population Nash-gap bound. Only the population optimal-response span is assumed bounded. The empirical and intermediate spans are bounded within the proof.

For a probability vector p, write $\mathrm{Var}_p(v) = \sum_s p(s)(v(s) - p^\top v)^2$. For a stochastic matrix Q, $\mathrm{Var}_Q(v)$ is the vector with s-th entry $\mathrm{Var}_{Q(s,\cdot)}(v)$. Vector inequalities, squares, and square roots are componentwise.

G.1 Oracle plug-in algorithm

Let

$$A_\times = \prod_{i=1}^{n} A_i, \qquad M = SA_\times, \qquad A_{\max} = \max_i A_i.$$

For a nominal row $p^0_{s,\boldsymbol{a}}$, define its support size

$$d(s,\boldsymbol{a}) = \left|\mathrm{supp}(p^0_{s,\boldsymbol{a}})\right|, \qquad d_{\max} = \max_{s,\boldsymbol{a}} d(s,\boldsymbol{a}) \le S.$$

Algorithm 1: Bernstein discounted reduction.

1. Input $\varepsilon \in (0,1]$, $\delta \in (0,1/2)$, a bound $H \ge H_{\mathrm{R}} \ge 1$, and an integer per-row sample budget $N \ge 1$ satisfying Theorem G.11. A supplied upper bound on $d_{\max}$ may be used to choose N.
2. Set
$$\gamma = 1 - \frac{\varepsilon}{2H}.$$
3. Draw N independent samples from every nominal row $p^0_{s,\boldsymbol{a}}$ and form the empirical center $\widehat{p}^0_{s,\boldsymbol{a}}$.
4. For each player, center the empirical TV ambiguity ball at $\widehat{p}^0_{s,\boldsymbol{a}}$ and retain the original radius $\sigma_{i,s,\boldsymbol{a}}$.
5. Return a measurable stationary product profile $\widehat{\pi}$ whose empirical unnormalized discounted Nash gap is at most $H/4$. An exact empirical Nash oracle is a special case.

The oracle condition means

$$\max_{i,s}\left\{\widehat{V}^{\star,\widehat{\pi}_{-i}}_{i,\gamma}(s) - \widehat{V}^{\widehat{\pi}}_{i,\gamma}(s)\right\} \le H/4.$$

These are discounted value gaps, not one-step Bellman residuals. The Nash oracle is charged no transition samples. Its computational cost is separate from the information-theoretic sample bound.

G.2 A single Bernstein event for every continuation vector

We begin with one nominal transition row. Let $p \in \Delta([S])$, let $X_1, \dots, X_N$ be independent with law p, and define

$$\widehat{p}(x) = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}\{X_k = x\}.$$

Write $d(p) = |\mathrm{supp}(p)|$.

Lemma G.1. *For every $\ell > 0$, with probability at least $1 - 2Se^{-\ell}$, the following two statements hold simultaneously:*

$$|\widehat{p}(x) - p(x)| \le \sqrt{\frac{2p(x)\ell}{N}} + \frac{2\ell}{3N} \qquad \textit{for every } x \in [S], \tag{61}$$

and

$$p(x) = 0 \quad \Longrightarrow \quad \widehat{p}(x) = 0. \tag{62}$$

Proof. Apply the bounded-variable Bernstein bound in (Maurer & Pontil, 2009, Theorem 3) to the independent Bernoulli variables $\mathbf{1}\{X_k = x\}$ and to their complements, with failure probability $e^{-\ell}$ for each tail. Their variance is $p(x)(1-p(x)) \le p(x)$, and the cited bound has linear term $\ell/(3N) \le 2\ell/(3N)$. A union bound over the S coordinates gives (61). If $p(x) = 0$, the coordinate is never sampled, almost surely, which gives (62) without any additional failure probability. □

The next lemma is the key game-level replacement for a fixed-vector Bernstein bound. Its probability event does not depend on the queried vector.

Lemma G.2. *On the event in Lemma G.1, every $v \in \mathbb{R}^S$ satisfies*

$$\left|(\widehat{p}-p)^\top v\right| \le \sqrt{\frac{2d(p)\ell}{N}\operatorname{Var}_p(v)} + \frac{2d(p)\ell}{3N}\operatorname{sp}(v)\,. \tag{63}$$

The inequality holds simultaneously over $\mathbb{R}^S$. In particular, v may be selected as an arbitrary measurable function of all transition samples.

Proof. This is a deterministic consequence of the coordinate event. Put $g = v - (p^\top v)\mathbf{1}$ and $d = d(p)$. The zero-sum property of $\widehat{p} - p$, support preservation, and Cauchy–Schwarz give

$$\begin{aligned}|(\widehat{p}-p)^\top v| &\le \sqrt{\frac{2\ell}{N}}\sum_{x\in\operatorname{supp}(p)}\sqrt{p(x)}|g(x)| + \frac{2\ell}{3N}\sum_{x\in\operatorname{supp}(p)}|g(x)| \\ &\le \sqrt{\frac{2d\ell}{N}\operatorname{Var}_p(v)} + \frac{2d\ell}{3N}\operatorname{sp}(v)\,,\end{aligned}$$

where $|g(x)| \le \operatorname{sp}(v)$. No event depends on v, establishing the stated simultaneous, and hence data-dependent, version. □

The span-localized discounted analysis also compares empirical and population conditional standard deviations.

Lemma G.3. *On the event in Lemma G.1, every $v \in \mathbb{R}^S$ satisfies the following inequalities with the numerical constant $C = 2$:*

$$\sqrt{\operatorname{Var}_{\widehat{p}}(v)} \le \sqrt{\operatorname{Var}_p(v)} + C\operatorname{sp}(v)\sqrt{\frac{d(p)\ell}{N}}, \tag{64}$$

$$\sqrt{\operatorname{Var}_p(v)} \le \sqrt{\operatorname{Var}_{\widehat{p}}(v)} + C\operatorname{sp}(v)\sqrt{\frac{d(p)\ell}{N}}. \tag{65}$$

Proof. Write $R = \operatorname{sp}(v)$ and $t = d(p)\ell/N$. For any c between $\min v$ and $\max v$, the vector $z = (v - c\mathbf{1})^2$ satisfies $\operatorname{sp}(z) \le R^2$ and $\operatorname{Var}_p(z) \le R^2 p^\top z$. With $c = p^\top v$, Lemma G.2 and the minimum-over-constants characterization of variance therefore give

$$\operatorname{Var}_{\widehat{p}}(v) \le \operatorname{Var}_p(v) + R\sqrt{2t\operatorname{Var}_p(v)} + \tfrac{2}{3}R^2 t \le \left(\sqrt{\operatorname{Var}_p(v)} + 2R\sqrt{t}\right)^2.$$

This proves (64). For the reverse direction, take $c = \widehat{p}^\top v$, $y = \sqrt{p^\top (v - c\mathbf{1})^2}$, and $\widehat{a} = \sqrt{\operatorname{Var}_{\widehat{p}}(v)}$. The same simultaneous bound gives

$$y^2 \le \widehat{a}^2 + R\sqrt{2t}\,y + \tfrac{2}{3}R^2 t.$$

Completing the square and taking the nonnegative root gives

$$y \le R\sqrt{t/2} + \sqrt{\widehat{a}^2 + \tfrac{7}{6}R^2 t} \le \widehat{a} + \left(\tfrac{1}{\sqrt{2}} + \sqrt{\tfrac{7}{6}}\right)R\sqrt{t} \le \widehat{a} + 2R\sqrt{t}.$$

Since $\sqrt{\operatorname{Var}_p(v)} \le y$, this proves (65). The argument is deterministic on the coordinate event; fixed-vector standard-deviation concentration alone would not give this uniform conclusion. □

G.3 TV CLIPPING AND ROBUST BELLMAN QUERIES

For $p \in \Delta([S])$, radius $\sigma \in [0,1]$, and vector v, define

$$\phi_{p,\sigma}(v) \triangleq \inf_{q \in \Delta([S]):\, \mathrm{TV}(q,p) \le \sigma} q^\top v.$$

The following standard TV-duality identity is the interface between the multinomial event and the robust Bellman operator.

Lemma G.4. *Let*

$$m = \min_x v(x), \qquad w = v - m\mathbf{1}, \qquad R = \mathrm{sp}(v)\,.$$

Then

$$\phi_{p,\sigma}(v) = m + \max_{\alpha \in [0,R]} \left\{ p^\top (w \wedge \alpha\mathbf{1}) - \sigma\alpha \right\}. \tag{66}$$

Proof. For $\sigma > 0$, apply (Shi et al., 2024b, Appendix A.3, Lemma 1) to the nonnegative vector $w = v - m\mathbf{1}$. That result uses the same normalization $\mathrm{TV}(q,p) = \frac{1}{2}\|q-p\|_1$. Since $\min w = 0$, its threshold interval is $[0,R]$ and $\min_x (w \wedge \alpha\mathbf{1})(x) = 0$, so its penalty is exactly $\sigma\alpha$. Translation by m proves (66). For $\sigma = 0$, both sides equal $p^\top v$ (choose $\alpha = R$ on the right). □

Corollary G.5. *Let $\mathcal{X}$ be finite, let $F \subseteq \mathcal{X}$ be a prescribed nonempty successor set, let $\sigma \in [0,1]$, and let $p \in \Delta(\mathcal{X})$ satisfy $\mathrm{supp}(p) \subseteq F$. Define the face-restricted ambiguity set*

$$\mathcal{Q}_F(p,\sigma) \triangleq \{q \in \Delta(\mathcal{X}) : \mathrm{supp}(q) \subseteq F,\ \mathrm{TV}(q,p) \le \sigma\}$$

and the corresponding support value

$$\phi_{p,\sigma,F}(v) \triangleq \inf_{q \in \mathcal{Q}_F(p,\sigma)} q^\top v.$$

For $v \in \mathbb{R}^{\mathcal{X}}$, set

$$m_F = \min_{x \in F} v(x), \qquad w_F(x) = v(x) - m_F \quad (x \in F), \qquad R_F = \max_{x \in F} v(x) - \min_{x \in F} v(x).$$

Then

$$\phi_{p,\sigma,F}(v) = m_F + \max_{\alpha \in [0,R_F]} \left\{ \sum_{x \in F} p(x) \min\{w_F(x), \alpha\} - \sigma\alpha \right\}. \tag{67}$$

Moreover, on the coordinatewise Bernstein event for the empirical row $\widehat{p}$, every possibly data-dependent v satisfies

$$|\phi_{\widehat{p},\sigma,F}(v) - \phi_{p,\sigma,F}(v)| \le \sqrt{\frac{2d(p)\ell}{N}\,\mathrm{Var}_p(v)} + \frac{2d(p)\ell}{3N} R_F. \tag{68}$$

Proof. Delete coordinates outside F. This identifies the prescribed face with $\Delta(F)$ and preserves total variation and expectations. Support preservation puts $\widehat{p}$ on that same face, so Lemma G.4 gives (67) for both centers. For $f_\alpha = w_F \wedge \alpha\mathbf{1}$, the difference of the two maxima is bounded by $\sup_{\alpha \in [0,R_F]} |(\widehat{p} - p)^\top f_\alpha|$. Moreover, $\mathrm{sp}(f)_\alpha \le R_F$ and, for independent $X, X' \sim p$,

$$\mathrm{Var}_p(f_\alpha) = \tfrac{1}{2}\mathbb{E}[(f_\alpha(X) - f_\alpha(X'))^2] \le \tfrac{1}{2}\mathbb{E}[(v(X) - v(X'))^2] = \mathrm{Var}_p(v).$$

Apply Lemma G.2 on F to obtain (68) simultaneously in α and v. The face is the prescribed F, including its nominally zero coordinates, and is not replaced by the empirical support. □

Lemma G.6. *On the event of Lemma G.1, for every $v \in \mathbb{R}^S$ and every $\sigma \in [0,1]$,*

$$|\phi_{\widehat{p},\sigma}(v) - \phi_{p,\sigma}(v)| \le \sqrt{\frac{2d(p)\ell}{N}\,\mathrm{Var}_p(v)} + \frac{2d(p)\ell}{3N}\,\mathrm{sp}(v)\,. \tag{69}$$

Proof. Take $F = [S]$ in Corollary G.5. Its deterministic proof applies to every $\sigma \in [0,1]$ on the same coordinate event. □

G.4 One event for the full game

Choose

$$\ell_0 = \log\left(\frac{2MS}{\delta}\right).$$

Apply Lemma G.1 to all $M = SA_\times$ primitive rows and take a union bound. With probability at least $1-\delta$, an event $\mathcal{E}_{\mathrm{row}}$ occurs on which Lemmas G.2, G.3, and G.6 hold simultaneously for every primitive row, every player-specific radius, every continuation vector, and every clipping threshold.

To pass from primitive rows to mixed actions, use nonnegative weights w_a with $\sum_a w_a = 1$. Then

$$\left|\sum_a w_a \xi_a\right| \le \sum_a w_a |\xi_a|,$$

and Jensen's inequality gives

$$\sum_a w_a \sqrt{z_a} \le \sqrt{\sum_a w_a z_a}.$$

Averaging these inequalities with any action probabilities introduces no action-cardinality factor. The same event will be used throughout the continuous center interpolation below. No additional union bound over players, opponent profiles, clipping thresholds, or interpolation parameters is needed.

G.5 Variance control along a segment of nominal centers

We prove the plug-in bound by continuously interpolating the nominal centers. The interpolation is an analysis device and requires no additional oracle calls or transition samples.

Work on $\mathcal{E}_{\mathrm{row}}$ and set

$$a = \frac{d_{\max}\ell_0}{N}, \qquad p_x(t) = (1-t)p_x + t\widehat{p}_x, \qquad 0 \le t \le 1.$$

Here x indexes a primitive uncertain row, and all player-specific radii satisfy $0 \le \sigma_{i,x} \le \sigma \le 1/2$, as in the model. For a row with prescribed successor face F_x, both centers and all ambiguity rows belong to $\Delta(F_x)$, and the same prescribed face is used along the entire path. The full-simplex case is $F_x = \mathcal{S}$. Exactly known rows have $\widehat{p}_x = p_x$. In this subsection $B = \mathrm{sp}(v)$ always denotes the global span, and therefore also bounds every face span. With a player fixed, we abbreviate $\sigma_{i,x}$ by σ_x.

Lemma G.7. *For every $t \in [0,1]$, vector v, and $q \in \mathcal{Q}_{F_x}(p_x(t), \sigma_x)$, the same event gives*

$$\sqrt{\mathrm{Var}_{p_x}(v)} \le \sqrt{\mathrm{Var}_q(v)} + B\sqrt{\sigma_x} + 3B\sqrt{a}, \tag{70}$$

$$\sup_{\alpha \in [0, B_x]} \left|(\widehat{p}_x - p_x)^\top [(v - m_x\mathbf{1}) \wedge \alpha\mathbf{1}]\right| \le \sqrt{2a}\sqrt{\mathrm{Var}_q(v)} + \sqrt{2a\sigma}\, B + 5aB, \tag{71}$$

where $m_x = \min_{y \in F_x} v(y)$ and $B_x = \max_{y \in F_x} v(y) - m_x$. The clipped vector is restricted to F_x.

Proof. First compare the original center to the interpolated center. Because $p_x(t) - p_x = t(\widehat{p}_x - p_x)$ and $0 \le t \le 1$, this difference satisfies the same coordinate bounds and support restriction as $\widehat{p}_x - p_x$. The deterministic proof of Lemma G.3, with $p_x(t)$ in place of $\widehat{p}_x$, therefore gives

$$\sqrt{\mathrm{Var}_{p_x}(v)} \le \sqrt{\mathrm{Var}_{p_x(t)}(v)} + 2B\sqrt{a}.$$

Next compare the interpolated center to the adverse row q. Use the standard minimum-over-constants argument (cf. Yang et al., 2026, Lemma 26). In the present normalization it gives

$$\mathrm{Var}_{p_x(t)}(v) \le \mathrm{Var}_q(v) + (p_x(t) - q)^\top (v - (q^\top v)\mathbf{1})^2 \le \mathrm{Var}_q(v) + \sigma_x B^2.$$

Here the squared vector has span at most B^2, which verifies the constant one in the TV term. Taking square roots and relaxing $2 \le 3$ proves (70). Lemma G.2, clipping's variance contraction, and (70) then give (71), using $3\sqrt{2} + 2/3 < 5$ and $\sigma_x \le \sigma$. These deductions hold for all t, v, q, and clipping thresholds on the same event. □

Lemma G.8 (Span-localized Bellman variance)**.** *Let Q be a stochastic matrix, $0 < \gamma < 1$, $r \in [0,1]^S$, and $v = r + \gamma Q v$. Write $\lambda = 1 - \gamma$, $B = \mathrm{sp}(v)$, and $\mathcal{R}_Q = (I - \gamma Q)^{-1}$. Then, componentwise,*

$$\gamma^2 \mathcal{R}_Q \operatorname{Var}_Q(v) \le \frac{2B}{\lambda}\mathbf{1}, \tag{72}$$

$$\gamma \left\| \mathcal{R}_Q \sqrt{\operatorname{Var}_Q(v)} \right\|_\infty \le \frac{\sqrt{2B}}{\lambda}. \tag{73}$$

Proof. We center the value before squaring the Bellman equation, so the final bound depends on its span rather than its absolute magnitude. The Neumann series for $\mathcal{R}_Q$ is entrywise nonnegative and gives $0 \le v \le \lambda^{-1}\mathbf{1}$ and $\mathcal{R}_Q\mathbf{1} = \lambda^{-1}\mathbf{1}$. Put $m = \min_s v(s)$, $u = v - m\mathbf{1}$, and $b = r - \lambda m\mathbf{1}$. Then $0 \le u \le B\mathbf{1}$, $b \le \mathbf{1}$, and $u = b + \gamma Q u$. Variance is invariant under additive constants, so

$$\begin{aligned}\gamma^2 \operatorname{Var}_Q(v) &= \gamma^2 Q u^2 - (u - b)^2 \\ &= \gamma^2 Q u^2 - u^2 + 2ub - b^2.\end{aligned}$$

Using $\gamma \mathcal{R}_Q Q = \mathcal{R}_Q - I$ and multiplying by $\mathcal{R}_Q$ gives the exact identity

$$\gamma^2 \mathcal{R}_Q \operatorname{Var}_Q(v) = -\lambda \mathcal{R}_Q u^2 - \gamma u^2 + \mathcal{R}_Q(2ub - b^2).$$

Since $u \ge 0$ and $b \le \mathbf{1}$, one has $2ub - b^2 \le 2u \le 2B\mathbf{1}$. The first two terms in the identity are nonpositive. The nonnegativity of $\mathcal{R}_Q$ therefore yields (72). Notice that no lower bound on b is used.

For each state s, the row $\nu_s = \lambda e_s^\top \mathcal{R}_Q$ is a probability vector. Jensen's inequality and (72) yield

$$\gamma e_s^\top \mathcal{R}_Q \sqrt{\operatorname{Var}_Q(v)} \le \frac{\gamma}{\lambda} \sqrt{\nu_s \operatorname{Var}_Q(v)} \le \frac{\sqrt{2B}}{\lambda}.$$

Taking the maximum over s proves (73). If $B = 0$, every conditional variance is zero, so both bounds still hold. □

G.6 Differentiating robust values along an affine path of centers

The following lemma permits a path of empirical models to be analyzed without assuming differentiability of an equilibrium oracle. Opponent policies remain fixed throughout the path, although they may have been chosen from the samples. Only the optimal-response value or the value of one fixed response is varied.

The proof separates two issues. First, finitely many affine LP bases represent all TV-minimizing rows along the center segment. Second, after refining the segment at changes of feasible bases, minimizers, and maximizing actions, the Bellman equation becomes a linear system with an affine transition matrix. The derivative formula follows by differentiating that system; no derivative of a Nash equilibrium correspondence is taken.

Lemma G.9. *Fix a discount $\gamma \in (0,1)$, player i, and stationary opponents θ. For each primitive row $x = (s, \boldsymbol{a})$, fix probability vectors $p_x, \widehat{p}_x$ on a nonempty prescribed face $F_x \subseteq \mathcal{S}$, and let $p_x(t) = p_x + t(\widehat{p}_x - p_x)$ for $t \in [0,1]$. The ambiguity row at t is*

$$\mathcal{Q}_x(t) = \{q \in \Delta(\mathcal{S}) : \mathrm{supp}(q) \subseteq F_x, \mathrm{TV}(q, p_x(t)) \le \sigma_{i,x}\}.$$

Rewards, radii, faces, and θ are fixed. Let v_t denote either (a) the robust discounted optimal-response value, or (b) the robust discounted value of a fixed stationary response μ_i. Assume rewards lie in $[0,1]$.

The map $t \mapsto v_t$ is globally Lipschitz. There is a finite partition of $[0,1]$ such that, on each open partition interval, v_t is a rational, continuously differentiable function of t. On each such interval, one can choose primitive minimizing rows $q_x(t) \in \mathcal{Q}_x(t)$ that are affine in t, and a stationary response μ_i that is constant on that interval, such that

$$v_t = r^{\mu_i,\theta} + \gamma Q_t v_t, \qquad Q_t(s,\cdot) = \sum_{a_i, \boldsymbol{a}_{-i}} \mu_i(a_i \mid s)\theta(\boldsymbol{a}_{-i} \mid s) q_{s,(a_i,\boldsymbol{a}_{-i})}(t).$$

In case (a), μ_i may be chosen deterministic and optimal on that interval. In case (b), it is the specified response. Each selected primitive row minimizes its expectation of v_t.

For $v \in \mathbb{R}^S$, define

$$m_x(v) = \min_{y \in F_x} v(y), \qquad B_x(v) = \max_{y \in F_x} v(y) - m_x(v),$$

$$L_x(v) = \sup_{0 \le \alpha \le B_x(v)} \left| \sum_{y \in F_x} (\widehat{p}_x(y) - p_x(y)) \min\{v(y) - m_x(v), \alpha\} \right|.$$

For every t in an open partition interval,

$$\dot{v}_t = \gamma(I - \gamma Q_t)^{-1} e_t, \qquad e_t(s) = \sum_{a_i, \boldsymbol{a}_{-i}} \mu_i(a_i \mid s)\theta(\boldsymbol{a}_{-i} \mid s)\dot{q}_{s,(a_i,\boldsymbol{a}_{-i})}(t)^\top v_t, \tag{74}$$

and every primitive derivative satisfies

$$\left|\dot{q}_x(t)^\top v_t\right| \le L_x(v_t). \tag{75}$$

Proof. **Step 1: continuity and a global Lipschitz bound.** For a fixed vector v, the TV clipping identity on the face F_x gives

$$\phi_x(t, v) = m_x(v) + \max_{0 \le \alpha \le B_x(v)} \left\{ \sum_{y \in F_x} p_x(t, y) \min\{v(y) - m_x(v), \alpha\} - \sigma_{i,x}\alpha \right\}.$$

The difference of two maxima is bounded in absolute value by the maximum of the absolute differences of their summands. Hence

$$|\phi_x(t, v) - \phi_x(u, v)| \le |t - u| L_x(v) \le |t - u| \operatorname{TV}(p_x, \widehat{p}_x) \operatorname{sp}(v). \tag{76}$$

For the last inequality, the clipped vector has span at most $\operatorname{sp}(v)$, and any zero-sum signed measure d satisfies $|d^\top f| \le \frac{1}{2}\|d\|_1 \operatorname{sp}(f)$. The inequality remains valid after taking opponent averages, response averages, or a maximum over own actions. Write T_t for the appropriate optimality or evaluation operator and put $D = \max_x \operatorname{TV}(p_x, \widehat{p}_x)$ and $\lambda = 1 - \gamma$. The fixed points satisfy $0 \le v_t \le \lambda^{-1}\mathbf{1}$. Using $v_t = T_t v_t$, the γ-contraction property, and (76), we obtain

$$\begin{aligned} \|v_t - v_u\|_\infty &\le \|T_t v_t - T_t v_u\|_\infty + \|T_t v_u - T_u v_u\|_\infty \\ &\le \gamma \|v_t - v_u\|_\infty + \gamma D |t - u| \operatorname{sp}(v_u) \\ &\le \gamma \|v_t - v_u\|_\infty + \frac{\gamma D}{\lambda} |t - u|. \end{aligned}$$

Subtracting the first term on the last line and dividing by $\lambda > 0$ proves $\|v_t - v_u\|_\infty \le \gamma D |t - u| / \lambda^2$. In particular, the path is continuous at all endpoints and is absolutely continuous, so its derivative bounds can be integrated.

Step 2: a finite collection of affine minimizing-row candidates. Identify a row's face with $\Delta(F_x)$ and introduce an auxiliary vector $z \in \mathbb{R}^{F_x}$. Its ambiguity set is the projection onto q of

$$\left\{(q, z) : q \ge 0, \quad \mathbf{1}^\top q = 1, \quad z \ge 0, \quad q - p_x(t) \le z, \quad p_x(t) - q \le z, \quad \mathbf{1}^\top z \le 2\sigma_{i,x}\right\}.$$

This polytope is nonempty because $(p_x(t), 0)$ is feasible, and it is bounded because $q \in \Delta(F_x)$ and $0 \le z(y) \le 2\sigma_{i,x}$. These facts include radius zero and singleton faces. Its constraint matrix does not depend on t, and its right-hand side is affine in t. At every vertex, the equality constraints together with the active inequality constraints have full rank in the ambient (q, z) coordinates: otherwise a nonzero null direction and sufficiently small displacements of both signs would contradict extremality. Choose a full-rank subcollection of these constraints. There are finitely many subcollections, each giving an affine candidate $(q_b(t), z_b(t))$ by a linear system with constant invertible coefficient matrix. This argument also covers lower-dimensional and degenerate polytopes; no nondegeneracy assumption is made.

Feasibility of each candidate is determined by finitely many affine inequalities and equalities in t. The zeroes of the nonidentically zero affine expressions, together with 0 and 1, partition $[0, 1]$ into finitely many intervals on which the list of feasible candidates is fixed. An identically zero expression imposes a permanent equality and adds no breakpoint. A linear objective over this nonempty compact polytope attains its minimum at a vertex. Thus, on each open interval and for every v, the robust

expectation is the minimum of the finitely many feasible affine-row expectations $q_b(t)^\top v$. Perform this construction for every primitive row and take the common finite partition. Behavior at the finitely many partition points will be handled by the global continuity already proved.

Step 3: finitely many rational value branches. On one interval from Step 2, select a feasible vertex candidate for every primitive row. In case (a), also select one own action at every state. In case (b), use the given response probabilities. This selection produces an affine stochastic matrix $Q(t)$ and a constant reward vector r. The candidate value is

$$w(t) = (I - \gamma Q(t))^{-1} r.$$

The inverse exists throughout the interval: $Q(t)$ is stochastic, its spectral radius is at most one, and $\gamma < 1$. Cramer's rule makes every coordinate of $w(t)$ rational in t. Call a selection valid when every selected primitive row minimizes its expectation of $w(t)$ and, in case (a), the selected own action at each state maximizes the corresponding response Bellman expression. More explicitly, the row conditions compare $q_x(t)^\top w(t)$ with $q_{x,b}(t)^\top w(t)$ for every feasible candidate b. Once all these row conditions hold, each own-action Bellman expression is computed using the selected rows for that action, so the maximizing-action conditions are also rational inequalities.

The denominator $\det(I - \gamma Q(t))$ is continuous and nonzero throughout the interval and hence has constant sign there. Multiplying by denominators with their correct signs reduces the validity conditions to finitely many polynomial inequalities. Each nonzero polynomial has finitely many roots; an identically zero polynomial creates a persistent tie and adds no breakpoint. There are finitely many selections. A further finite partition therefore makes the validity of every selection constant on each open interval. No uniqueness of optimal actions or minimizing rows is required.

At every parameter value in an open interval, the robust Bellman equation admits primitive minimizing vertices of the lifted polytopes and, in case (a), maximizing deterministic own actions. Choose them at that value. Their Bellman equality implies that the corresponding candidate $w(t)$ equals v_t, so at least one selection is valid there. On each final open interval choose a selection valid at any one interior point; by the preceding partition it is valid throughout that interval. Its candidate is a fixed point of T_t, so contraction uniqueness gives $w(t) = v_t$ throughout the interval. This proves the asserted rational smoothness and the existence of constant response choices and affine minimizing rows there.

Step 4: controlling the derivative of a selected minimizing row. Fix an interior point t of one final partition interval and set $v = v_t$. Keep this vector fixed while the center parameter changes. The selected candidate $q_x(u)$ remains feasible for u on both sides of t within the interval. Consequently,

$$\phi_x(u, v) \le q_x(u)^\top v, \qquad \phi_x(t, v) = q_x(t)^\top v.$$

Equation (76) also gives $\phi_x(u, v) \ge \phi_x(t, v) - L_x(v)|u - t|$. Combining these relations and using the affine form of q_x yields

$$(u - t)\dot{q}_x(t)^\top v = q_x(u)^\top v - q_x(t)^\top v \ge -L_x(v)|u - t|.$$

For $u > t$, division by $u - t > 0$ gives $\dot{q}_x(t)^\top v \ge -L_x(v)$. For $u < t$, division by $u - t < 0$ reverses the inequality and gives $\dot{q}_x(t)^\top v \le L_x(v)$. This proves (75); it does not require that $q_x(u)$ minimize the expectation of the fixed vector v_t away from t.

Step 5: differentiating the Bellman equality. On a final open interval, r is constant, Q_t is affine, and v_t is continuously differentiable. The product rule applied to $v_t = r + \gamma Q_t v_t$ gives

$$\dot{v}_t = \gamma \dot{Q}_t v_t + \gamma Q_t \dot{v}_t.$$

The definition of Q_t and the fact that the response and opponent weights are constant show that $\dot{Q}_t v_t = e_t$. Multiplying $(I - \gamma Q_t)\dot{v}_t = \gamma e_t$ by the existing inverse proves (74). □

G.7 Simultaneous response guarantee

Theorem G.10. *Fix any player and any stationary opponent profile in the sampling model above, with known rewards in $[0, 1]$ and known radii bounded by the common number σ. Allow either full TV balls or TV balls restricted to prescribed nonempty successor faces containing the respective nominal supports. The same radii and faces are retained at the empirical centers. Let $v^\star$ and $\widehat{v}^\star$ denote the*

population and empirical optimal-response values at discount $\gamma \in (0,1)$, and let v^μ and $\widehat{v}^\mu$ denote the corresponding values of a stationary response μ. Suppose $R_0 \geq \max\{1, \mathrm{sp}(v^\star)\}$. Set $\lambda = 1-\gamma$ and choose $e \in (0, R_0]$. On $\mathcal{E}_{\mathrm{row}}$, every stationary response $\widehat{\mu}$ satisfying

$$0 \leq \widehat{v}^\star - \widehat{v}^{\widehat{\mu}} \leq \varepsilon_{\mathrm{opt}}\mathbf{1}, \qquad \varepsilon_{\mathrm{opt}} \leq e/4,$$

obeys $\|v^\star - v^{\widehat{\mu}}\|_\infty \leq e$, provided

$$N \geq C d_{\max}\ell_0 \left[\frac{R_0 + \sigma R_0^2}{\lambda^2 e^2} + \frac{R_0}{\lambda e}\right] \tag{77}$$

for a universal constant C. The conclusion is simultaneous over all players, opponent profiles, and responses satisfying these hypotheses, including choices made from the same transition data. No bound on a population candidate-policy span or an empirical optimal span is assumed.

Proof. **Derivative control.** Set $a = d_{\max}\ell_0/N$. Fix the opponent profile and consider either the optimal-response path or the evaluation path of one fixed response. At every point in an open interval of its finite partition, and hence almost everywhere, let $Q(t)$ and $e(t)$ be the stochastic matrix and perturbation vector from Lemma G.9. They satisfy

$$v(t) = r(t) + \gamma Q(t)v(t), \qquad v'(t) = \gamma(I - \gamma Q(t))^{-1}e(t), \qquad 0 \leq r(t) \leq \mathbf{1}.$$

Write $B = \mathrm{sp}(v(t))$. For each primitive row used by $Q(t)$, combine (75) with Lemma G.7 to bound the absolute center derivative by $\sqrt{2a}\sqrt{\mathrm{Var}_q(v(t))} + \sqrt{2a\sigma}B + 5aB$. Average with the current response and opponent action probabilities. The triangle inequality, Jensen's inequality, and the law of total variance give

$$|e(t)| \leq \sqrt{2a}\sqrt{\mathrm{Var}_{Q(t)}(v(t))} + (\sqrt{2a\sigma}B + 5aB)\mathbf{1}.$$

In detail, if a row of Q is $\sum_j w_j q_j$, then

$$\sum_j w_j \sqrt{\mathrm{Var}_{q_j}(v)} \leq \sqrt{\sum_j w_j \,\mathrm{Var}_{q_j}(v)} \leq \sqrt{\mathrm{Var}_{\sum_j w_j q_j}(v)}.$$

Thus the averaging introduces no action-count factor. Applying Lemma G.8 and the row sums of the resolvent yields

$$\|v'(t)\|_\infty \leq K(B) := \frac{2\sqrt{aB} + \sqrt{2a\sigma}\,B + 5aB}{\lambda}. \tag{78}$$

This calculation uses the adverse rows for the current value and current center. Their conditional variance is therefore controlled by their own Bellman equation.

A common localization radius. We use the same radius $4R_0$ for both paths. The function K is nondecreasing on $[0,\infty)$. Choosing $C = 2^{18} = 262144$ in (77) ensures

$$K(4R_0) \leq e/4. \tag{79}$$

Indeed, put $L = R_0 + \sigma R_0^2$. The sample condition gives $a \leq \lambda^2 e^2/(CL)$ and $a \leq \lambda e/(CR_0)$. Therefore

$$\frac{K(4R_0)}{e} \leq \frac{4+4\sqrt{2}}{\sqrt{C}} + \frac{20}{C} = \frac{4+4\sqrt{2}}{512} + \frac{20}{262144} < 0.019 < \frac{1}{4}.$$

Step 1: control the optimal-response path from population to empirical centers. Let $v^\star(t)$ be this path. It is continuous and absolutely continuous by Lemma G.9. Its initial span is at most R_0. If its span first reached $4R_0$ at time $\tau \leq 1$, then for $0 \leq t \leq \tau$, integrating (78) and (79) would give

$$\|v^\star(t) - v^\star(0)\|_\infty \leq te/4 \leq e/4.$$

The span triangle inequality would then imply $\mathrm{sp}(v^\star(\tau)) \leq R_0 + e/2 \leq 3R_0/2$, a contradiction. There is no such first crossing. The same integral estimate therefore holds throughout $[0,1]$, and in particular

$$\|\widehat{v}^\star - v^\star\|_\infty \leq e/4, \qquad \mathrm{sp}(\widehat{v}^\star) \leq R_0 + e/2. \tag{80}$$

The factor $e/2$ in the span estimate comes from $\mathrm{sp}(x-y) \le 2\|x-y\|_\infty$. Thus the $e/4$ uniform-norm path bound is exactly what is needed for the first-crossing contradiction.

Step 2: control the selected response path from empirical to population centers. Fix the response $\widehat{\mu}$ and keep it fixed along the entire path. This is a deterministic choice after the simultaneous event has been fixed. Its empirical value satisfies

$$\mathrm{sp}\left(\widehat{v}^{\widehat{\mu}}\right) \le \mathrm{sp}(\widehat{v}^\star) + \varepsilon_{\mathrm{opt}} \le R_0 + 3e/4 \le 7R_0/4.$$

Here the solver-error vector has coordinates in $[0, \varepsilon_{\mathrm{opt}}]$, so its span is at most $\varepsilon_{\mathrm{opt}}$. To make the backward argument explicit, set $w(t) = v^{\widehat{\mu}}(1-t)$ for $t \in [0,1]$, where the argument of $v^{\widehat{\mu}}$ denotes its center parameter. Absolute continuity and (78) give $\|w'(t)\|_\infty \le K(\mathrm{sp}(w(t)))$ almost everywhere, because reversing the parameter only changes the sign of the derivative. Starting from $\mathrm{sp}(w(0)) \le 7R_0/4$, if the span first reached $4R_0$ at some $t = \tau$, integration up to that time would give $\|w(\tau) - w(0)\|_\infty \le e/4$. Its span would then be at most $R_0 + 3e/4 + e/2 \le 9R_0/4$, again a contradiction. Thus the same integral bound holds on all of $[0,1]$, and

$$\|v^{\widehat{\mu}} - \widehat{v}^{\widehat{\mu}}\|_\infty \le e/4. \tag{81}$$

Step 3: combine the two comparisons. Population optimality gives $v^\star - v^{\widehat{\mu}} \ge 0$ componentwise. Insert the two empirical values and apply the triangle inequality:

$$\begin{aligned}\|v^\star - v^{\widehat{\mu}}\|_\infty &\le \|v^\star - \widehat{v}^\star\|_\infty + \|\widehat{v}^\star - \widehat{v}^{\widehat{\mu}}\|_\infty + \|\widehat{v}^{\widehat{\mu}} - v^{\widehat{\mu}}\|_\infty \\ &\le e/4 + \varepsilon_{\mathrm{opt}} + e/4 \le 3e/4 \le e.\end{aligned}$$

All inequalities were deterministic consequences of one event and the stated span bound. They apply to every fixed choice of opponents and response, with the same constants. This proves the asserted uniformity without conditioning on an empirical opponent profile. □

Relation to the robust-MDP plug-in literature. The leading terms $R_0 + \sigma R_0^2$ also appear in the single-agent bound of (Yang et al., 2026, Corollary 3). That result concerns one fixed robust MDP and uses a different concentration argument and solver tolerance. Here the coordinate event and center-path proof establish the additional uniformity over opponent profiles, including those selected from the same data, while allowing prescribed successor faces and a value-gap tolerance $e/4$. A fixed-response robust-MDP theorem cannot be applied by conditioning on an opponent profile chosen from the same transition samples: that conditioning can change the sampling law.

G.8 End-to-end sample complexity

Theorem G.11. *Consider the finite player-specific game with known rewards in $[0,1]$, primitive rectangular TV radii at most $\sigma \le 1/2$, and independent nominal generative samples. Let $H \ge H_{\mathrm{R}} \ge 1$, $0 < \varepsilon \le 1$, and $0 < \delta < 1/2$. Run Algorithm 1 with any measurable oracle satisfying its empirical discounted Nash-gap condition. There is a universal constant C such that*

$$N \ge C d_{\max} \frac{H + \sigma H^2}{\varepsilon^2} \log\left(\frac{2S^2 A_\times}{\delta}\right) \tag{82}$$

implies $\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \le \varepsilon$ with probability at least $1-\delta$, simultaneously for every oracle output with empirical unnormalized discounted Nash gap at most $H/4$. In particular, the conclusion holds for every exact empirical Nash oracle. Taking N to be the ceiling of the right side yields

$$N_{\mathrm{tot}} \le C' S A_\times d_{\max} \frac{H + \sigma H^2}{\varepsilon^2} \log\left(\frac{2S^2 A_\times}{\delta}\right) \tag{83}$$

for another universal constant C'. If $d_{\max}$ is not known, a supplied upper bound may replace it, and S always suffices.

Proof. Theorem D.3 supplies an exact empirical equilibrium, so the oracle requirement is feasible. Set $\gamma = 1 - \varepsilon/(2H)$, giving $\lambda = \varepsilon/(2H)$ and $\gamma \ge 1/2$. On $\mathcal{E}_{\mathrm{row}}$, fix any output $\widehat{\pi}$ satisfying the empirical gap condition, and any player i. Put $\theta = \widehat{\pi}_{-i}$. Theorem C.9 bounds the population

optimal-response span by $R_0 = 2H$. The oracle condition gives $0 \le \widehat{V}^{\star,\theta}_{i,\gamma} - \widehat{V}^{\widehat{\pi}}_{i,\gamma} \le (H/4)\mathbf{1}$, which is precisely the solver condition of Theorem G.10 with $e = H$.

Apply Theorem G.10 with $e = H$. Its sample bracket satisfies

$$\frac{R_0 + \sigma R_0^2}{\lambda^2 e^2} + \frac{R_0}{\lambda e} = \frac{8H + 16\sigma H^2}{\varepsilon^2} + \frac{4H}{\varepsilon} \le 20\frac{H + \sigma H^2}{\varepsilon^2},$$

where $\varepsilon \le 1$. The event uses $\ell_0 = \log(2S^2A_\times/\delta)$, so (82) implies this sample condition after increasing C. Therefore the population discounted Nash gap is at most H for every player simultaneously. The reduction gives

$$\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \le \lambda[H_{\mathrm{R}} + \mathrm{Gap}_\gamma(\widehat{\pi})] \le \frac{\varepsilon}{2H}(H + H) = \varepsilon.$$

The row event has probability at least $1 - \delta$. Its deterministic uniformity over opponent profiles and selected responses proves the oracle assertion. Finally, choose N as the indicated ceiling and multiply by $SA_\times$; the extra ceiling term is absorbed by C' because $H \ge 1$, $d_{\max} \ge 1$, and $\varepsilon \le 1$. □

Corollary G.12 (Dense tabular transitions)**.** *Since $d_{\max} \le S$, the total sample complexity is*

$$\widetilde{O}\left(S^2A_\times \frac{H + \sigma H^2}{\varepsilon^2}\right).$$

Corollary G.13 (Bounded-support transitions)**.** *If $d_{\max} = O(1)$, the total sample complexity is*

$$\widetilde{O}\left(SA_\times \frac{H + \sigma H^2}{\varepsilon^2}\right).$$

H LOWER BOUNDS

We distinguish adaptive sampling from balanced empirical-model learning whose guarantee must hold for every exact empirical Nash oracle. The first quantifier concerns every learning algorithm. The second requires that a statistical guarantee survive an adverse, but exact, choice of empirical equilibrium. We keep these experiments separate throughout. The robust-control dependence $H + \sigma H^2$ already appears in (Yang et al., 2026, Theorem 1). Its balanced per-row experiment does not directly imply the adaptive expected-budget claim below; we retain that embedding and its information argument. The oracle constructions additionally require realizable, sample-dependent equilibrium continuations.

H.1 ADAPTIVE GENERATIVE-MODEL LOWER BOUND

This subsection proves Theorem 6.1. We place one unknown best action in each two-state component. Statewise approximate optimality identifies every component, and the following lemma lower-bounds the observations spent in one component even when queries are interleaved with observations elsewhere.

H.1.1 AN ADAPTIVE IDENTIFICATION LEMMA

Lemma H.1. *Let $A \ge 2$, $0 < b < b + \Delta < 1$, and*

$$k_0 = \mathrm{KL}\big(\mathrm{Ber}(b) \,\|\, \mathrm{Ber}(b + \Delta)\big) > 0.$$

Under hypothesis $a \in [A]$, arm a has mean $b + \Delta$ and every other arm has mean b. Suppose an adaptive procedure returns the correct special-arm label with probability at least $3/4$ under each hypothesis. Nontermination is allowed on its failure event. Let $T \in \mathbb{N} \cup \{0, \infty\}$ count only observations of these A arms over its entire execution, up to return if it returns. The procedure may also use auxiliary randomness and observation streams whose conditional sampling laws, given the past and the chosen auxiliary query, do not depend on a. Then

$$\frac{1}{A}\sum_{a=1}^{A} \mathbb{E}_a[T] \ge \frac{A}{256k_0}. \tag{84}$$

In particular the same lower bound holds with $\kappa = \max\{k_0, \mathrm{KL}(\mathrm{Ber}(b+\Delta)\|\mathrm{Ber}(b))\}$ *in place of* k_0.

Proof. Use the finite-transcript change-of-measure identity of (Kaufmann et al., 2016, Appendix A.1, equation (18)). We retain the truncation needed here because termination is required only on correct-output paths and auxiliary observations are not charged to T. Let $\bar{t} = A^{-1}\sum_a \mathbb{E}_a T$. If $\bar{t} = \infty$, there is nothing to prove. If $\bar{t} = 0$, no informative observation is ever collected under any alternative. The output law is then independent of the special arm, so its average success probability is at most $1/A \le 1/2$, a contradiction. Thus $0 < \bar{t} < \infty$. Put $n = \lfloor 8\bar{t} \rfloor$. Stop before informative query $n+1$, and also after L total operations, returning an arbitrary label at either cutoff. For this finite procedure, denote the output by $\widehat{a}_L$. Because T is integer valued and $n+1 > 8\bar{t}$, Markov's inequality gives $A^{-1}\sum_a \mathbb{P}_a(T > n) \le \bar{t}/(n+1) \le 1/8$. Every correct-output path with $T \le n$ survives the second cutoff for all sufficiently large L, so Fatou's lemma gives

$$\liminf_{L\to\infty} \frac{1}{A}\sum_a \mathbb{P}_a(\widehat{a}_L = a) \ge \frac{5}{8}. \tag{85}$$

Under the null model all arms have mean b. The Bernoulli laws are mutually absolutely continuous; query rules and auxiliary conditional laws coincide under the null and alternatives. The query choices themselves therefore add no conditional relative entropy. Neither do the auxiliary observations, whose conditional kernels agree. Only an observation of arm a contributes to the cited likelihood identity. Let $\mathbb{P}_{0,L}, \mathbb{P}_{a,L}$ be the twice-truncated transcript laws and N_a its arm-a observation count. The identity yields

$$\mathrm{KL}(\mathbb{P}_{0,L}\|\mathbb{P}_{a,L}) = k_0 \mathbb{E}_{0,L}[N_a]. \tag{86}$$

Pinsker's inequality, Cauchy–Schwarz, and $\sum_a N_a \le n$ imply

$$\frac{1}{A}\sum_a \mathbb{P}_a(\widehat{a}_L = a) \le \frac{1}{A} + \frac{1}{A}\sum_a \sqrt{k_0 \mathbb{E}_{0,L} N_a/2} \le \frac{1}{A} + \sqrt{\frac{k_0 n}{2A}}.$$

Let $L \to \infty$. Since $A \ge 2$, the last square root is at least $1/8$. Hence $n \ge A/(32k_0)$ and $n \le 8\bar{t}$ proves (84). No guarantee at the null model is used. □

H.1.2 Adaptive robust-game lower bound

Theorem H.2. *There are universal constants* $c, c_0 > 0$ *such that the following holds. Let* $S \ge 4$, $A_\times \ge 2$, $H \ge 4$, $0 \le \sigma \le 1/2$, *and* $0 < \varepsilon \le c_0$. *There is a class of player-specific state–joint-action rectangular TV robust Markov games with known rewards in* $[0,1]$, *nominal support size at most two, and* $H_{\mathrm{R}} \le H$, *such that every adaptive generative-model learner which returns a stationary* ε*-robust average-reward Nash equilibrium with probability at least* $3/4$ *on every instance satisfies*

$$\sup_{\mathcal{M}} \mathbb{E}_{\mathcal{M}}[T] \ge cSA_\times \frac{H + \sigma H^2}{\varepsilon^2}. \tag{87}$$

Here T *is its total number of nominal transition observations. The class has one active player with* $A_\times$ *actions and any number of additional players with one action and zero reward. Consequently this is a lower bound for classes specified by the joint-action budget while allowing its allocation across players to vary. For an arbitrary prescribed vector of action cardinalities, this embedding alone establishes the same bound with* $\max_i A_i$ *in place of* $A_\times$.

Proof. **1. Construct independent nominal components.** Let $m = \lfloor S/2 \rfloor$ and introduce states $\{d_j, r_j : j \in [m]\}$. A possible extra state for odd S is specified below. Give the active player $A = A_\times$ actions. Set

$$q = \frac{1}{H}, \qquad b = \sigma + \frac{1}{H}, \qquad \Delta_\varepsilon = \frac{10\varepsilon}{H}, \qquad c_0 = 10^{-2}.$$

The hidden parameter is $\theta = (\theta_1, \dots, \theta_m) \in [A]^m$. At r_j, all actions give reward one and the known row

$$P^0(d_j \mid r_j) = q, \qquad P^0(r_j \mid r_j) = 1 - q,$$

with ambiguity radius zero. At d_j, all actions give reward zero and

$$P^0(r_j \mid d_j, a) = b + \Delta_\varepsilon \mathbf{1}\{a = \theta_j\}, \qquad P^0(d_j \mid d_j, a) = 1 - P^0(r_j \mid d_j, a), \tag{88}$$

with ambiguity radius σ. These are full TV balls on $\Delta(\mathcal{S})$, with no restriction to the nominal two-state support. Since $b \le 3/4$ and $\Delta_\varepsilon \le 1/40$, all displayed probabilities lie strictly between zero and one. Every nominal row has at most two successors.

2. Verify a common optimal gain despite cross-component ambiguity. Put

$$p_\star = q + \Delta_\varepsilon, \qquad \rho_\star = \frac{p_\star}{p_\star + q}, \qquad B = \frac{1}{p_\star + q}, \qquad h(d_j) = 0, \quad h(r_j) = B.$$

Every decision state has bias zero and every reward state has bias B. For any nominal decision row with probability u of r_j, the minimum expectation of h over its full TV ball is $(u - \sigma)B$. Indeed, writing $\mathcal{R} = \{r_1, \dots, r_m\}$, $\mathrm{TV}(p, P^0) \le \sigma$ implies $p(\mathcal{R}) \ge P^0(\mathcal{R}) - \sigma = u - \sigma$. Hence $p^\top h \ge (u - \sigma)B$. Equality is attained by removing σ mass from r_j and placing it at d_j. Allowing nature to send mass to other components cannot improve this minimum, because all their bias values also lie in $[0, B]$.

The special action therefore gives robust continuation $p_\star B = \rho_\star$, while every other action gives $qB < \rho_\star$. At r_j, the Bellman right-hand side equals

$$1 + qh(d_j) + (1 - q)h(r_j) = 1 + (1 - q)B = \frac{p_\star + 1}{p_\star + q} = \rho_\star + h(r_j).$$

Thus $(\rho_\star, h)$ solves the constant-gain optimality equation. For odd S, add a state e with known reward $\rho_\star$, zero ambiguity radius, deterministic transition to d_1, and $h(e) = 0$. Its Bellman equation also holds. The reward $\rho_\star$ depends only on H and ε, not on any hidden action label, and therefore reveals no information about θ.

Proposition C.6 identifies $\rho_\star$ as the robust optimal average reward from every initial state, against stationary or history-dependent nature. Its maximizing action is the special action in every component. Moreover,

$$\mathrm{sp}(h) = B = \frac{1}{2/H + \Delta_\varepsilon} \le H/2 \le H.$$

Each inactive player's zero reward admits the solution $(0, 0)$ against every opponent profile. Thus the game has $H_{\mathrm{R}} \le H$.

3. Show that approximate equilibrium identifies every component. For any output stationary policy, write $w_j = \pi_1(\theta_j \mid d_j)$. Nature can use the within-component row that removes σ mass from r_j and adds it at d_j for every action. Under this particular admissible stationary kernel, the chain started at d_j stays in $\{d_j, r_j\}$ and has transition probabilities $p(w_j) = q + w_j\Delta_\varepsilon$ from d_j to r_j and q in the reverse direction. Both are positive. Its stationary probability of r_j, and hence its average reward, is

$$\rho(w_j) = \frac{p(w_j)}{p(w_j) + q}. \tag{89}$$

The robust payoff of the policy is at most this quantity. This upper bound does not assert that the chosen kernel is worst case for the output policy; one admissible kernel is enough.

If $w_j \le 1/2$, direct subtraction gives

$$\begin{aligned}
\rho_\star - \rho(w_j) &= \frac{p_\star(p(w_j) + q) - p(w_j)(p_\star + q)}{(p_\star + q)(p(w_j) + q)} \\
&= \frac{q(1 - w_j)\Delta_\varepsilon}{(p_\star + q)(p(w_j) + q)} \\
&\ge \frac{5\varepsilon/H^2}{(2.1/H)^2} > \varepsilon.
\end{aligned}$$

The denominators are positive. The inequality uses $1 - w_j \ge 1/2$ and $10\varepsilon \le 0.1$. Since the Nash gap is statewise, every ε-equilibrium must have $w_j > 1/2$ for every j. Its unique action of probability greater than $1/2$ at d_j is therefore θ_j. On unsuccessful outputs, use any deterministic label if there is no such action.

4. Bound the information in one transition observation. A query to (d_j, a) is a Bernoulli observation with mean b or $b + \Delta_\varepsilon$. We use the elementary Bernoulli bound $\mathrm{KL}(\mathrm{Ber}(u)\|\mathrm{Ber}(v)) \le (u-v)^2/[v(1-v)]$ for $u, v \in (0,1)$, obtained by applying $\log z \le z - 1$ to the two logarithms in the divergence. Both possible denominators are bounded below by $b/5$, because $b \le 3/4$ and $b + \Delta_\varepsilon \le 31/40$. Consequently, with κ the maximum of the two directed divergences,

$$\kappa \le \frac{5\Delta_\varepsilon^2}{b} = \frac{500\varepsilon^2}{H + \sigma H^2}. \tag{90}$$

5. Sum the adaptive identification costs. It suffices to consider a learner with finite worst-case expected sample count, since otherwise the claim already holds. Draw the hidden vector uniformly from $[A]^m$ and let T_j count its queries to decision rows in component j. Fix every coordinate θ_k with $k \ne j$. For each possible value of θ_j, the learner identifies θ_j with probability at least $3/4$, because successful equilibrium output identifies the entire hidden vector. Observations from all other rows can be generated as auxiliary observations whose conditional laws do not depend on θ_j. Although the learner's choices of those queries may depend on its past observations, their conditional observation kernels still carry zero likelihood-ratio increment for θ_j. Lemma H.1 therefore gives

$$\mathbb{E}[T_j \mid (\theta_k)_{k \ne j}] \ge c\frac{A}{\kappa} \ge c' A \frac{H + \sigma H^2}{\varepsilon^2},$$

where the expectation also averages over uniform θ_j. Average over the other coordinates and add these inequalities. Since $T \ge \sum_j T_j$ and $m = \lfloor S/2 \rfloor \ge S/3$ for $S \ge 4$,

$$\mathbb{E}[T] \ge c' m A \frac{H + \sigma H^2}{\varepsilon^2} \ge c'' S A \frac{H + \sigma H^2}{\varepsilon^2}.$$

This prior average is no larger than the worst-case expected count, which proves (87). □

Remark H.3 (The bounded range $1 \le H < 4$)**.** *The same order holds on this bounded range by a non-robust construction. Use $q = b = 1/2$, decision-row ambiguity radius zero, and special-arm increment 10ε. Its optimal bias span is $1/(1+10\varepsilon) \le 1$, and if $w \le 1/2$ its loss is at least $2.5\varepsilon/(1.1)^2 > \varepsilon$ for $\varepsilon \le 10^{-2}$. The Bernoulli divergence is $O(\varepsilon^2)$, so the preceding identification argument gives $\Omega(SA_\times/\varepsilon^2)$. All radii are at most the allowed σ, while $H + \sigma H^2 < 12$ on this parameter range.*

H.2 Every-oracle lower bound under balanced sampling

This subsection proves Theorem 6.2. The first construction isolates the successor-support obstruction under a common positive TV radius. The joint construction then supplies the support–span–ambiguity dependence and its all-budget completion.

H.2.1 A successor lower bound with strictly positive TV ambiguity

The successor factor persists when every primitive ambiguity radius is the same positive number. At each branch, a coordination game admits a low and a high equilibrium. The oracle uses the signs of the hub-row estimation errors to choose between them. A high-valued anchor keeps the robust continuation difference in the affine part of the TV trimming formula. The small-ball lemma quantifies this aligned fluctuation; the theorem then checks every empirical Nash inequality and transfers the hub error to an average-reward loss.

Definition H.4. *Draw N independent nominal observations per primitive state–joint-action row, independently across rows, form the empirical centers, retain the prescribed radii, and return an exact stationary discounted robust Nash equilibrium $\mathfrak{O}_\gamma(\widehat{P}^0)$. An oracle-uniform guarantee requires*

$$\mathbb{P}_{P^0}\Big(\mathrm{Gap}_{\mathrm{av}}(\mathfrak{O}_\gamma(\widehat{P}^0)) \le \varepsilon\Big) \ge 1 - \delta$$

for every game in the stated class and every measurable exact oracle. Rows with identical population centers are sampled and estimated separately. Here $\mathrm{TV}(p,q) = \frac{1}{2}\|p - q\|_1$, and $\phi_{p,\sigma}(v) = \min_{q:\mathrm{TV}(q,p)\le\sigma} q^\top v$.

Lemma H.5. *Let $m \geq 2$, $N \geq 1024m$, and let p be the distribution on $\{z, 1, \ldots, m\}$ given by*

$$p(z) = \frac{3}{4}, \qquad p(j) = \frac{1}{4m} \quad (1 \leq j \leq m).$$

Let $\widehat{p}_A, \widehat{p}_B$ be independent empirical distributions formed from N independent observations each, and put

$$\Delta = \widehat{p}_A - \widehat{p}_B, \qquad T = \sum_{j=1}^{m} (-\Delta(j))_+, \qquad k = \frac{1}{16384}.$$

Define

$$\mathcal{E} = \left\{ T \geq k\sqrt{\frac{m}{N}},\ \widehat{p}_A(z) \geq \frac{1}{2},\ \widehat{p}_B(z) \geq \frac{1}{2} \right\}.$$

Then

$$\Pr(\mathcal{E}) \geq \max\left\{ \frac{1}{3},\ 1 - \exp\left(-\frac{m}{1024}\right) \right\}. \tag{91}$$

In particular, the anchor has empirical mass at least every $\sigma \in (0, 1/2]$ on the same event.

Proof. We first lower-bound the mean fluctuation, then prove its concentration for large m. A one-coordinate anti-concentration estimate supplies the constant success probability when m is small. Finally we enforce the anchor condition.

Write X_j and Y_j for the counts in coordinate j from the two samples, and set $Z_j = X_j - Y_j = N\Delta(j)$. For a fixed j let $q = 1/(4m)$ and

$$\tau^2 = \mathbb{E}Z_j^2 = 2Nq(1-q) \geq \frac{7N}{16m} \geq 448.$$

The distribution of Z_j is symmetric about zero. It is a sum of $2N$ independent centered random variables bounded in absolute value by one. For its independent centered summands U_ℓ, the fourth-moment expansion and $\mathbb{E}U_\ell^4 \leq \mathbb{E}U_\ell^2$ give

$$\mathbb{E}Z_j^4 = \sum_{\ell} \mathbb{E}U_\ell^4 + 6\sum_{\ell<k} \mathbb{E}U_\ell^2 \mathbb{E}U_k^2 \leq \tau^2 + 3\tau^4 \leq 4\tau^4,$$

where the last inequality uses $\tau^2 \geq 448 > 1$. Hölder's inequality implies

$$\mathbb{E}Z_j^2 \leq (\mathbb{E}|Z_j|)^{2/3} (\mathbb{E}Z_j^4)^{1/3}.$$

Rearranging gives $\mathbb{E}|Z_j| \geq (\mathbb{E}Z_j^2)^{3/2}/(\mathbb{E}Z_j^4)^{1/2} \geq \tau/2$. By symmetry, $\mathbb{E}(-Z_j)_+ = \mathbb{E}|Z_j|/2$, and hence

$$\mathbb{E}T = \frac{1}{2N}\sum_{j=1}^{m} \mathbb{E}|Z_j| \geq \frac{m\tau}{4N} \geq \frac{\sqrt{7}}{16}\sqrt{\frac{m}{N}}. \tag{92}$$

Independence across the coordinates is not used here.

For the concentration calculation, observe that

$$T = \max_{D \subseteq \{1,\ldots,m\}} \big(\widehat{p}_B(D) - \widehat{p}_A(D)\big).$$

Replacing one observation changes every quantity inside this maximum by at most $1/N$, and hence changes T by at most $1/N$. McDiarmid's bounded-differences inequality, applied to the $2N$ independent observations, therefore gives

$$\Pr\left(T < \tfrac{1}{2}\mathbb{E}T\right) \leq \exp\left(-\frac{N(\mathbb{E}T)^2}{4}\right) \leq \exp\left(-\frac{7m}{1024}\right).$$

Since $k \leq \sqrt{7}/32$, it follows that

$$\Pr\left(T < k\sqrt{\frac{m}{N}}\right) \leq \exp\left(-\frac{7m}{1024}\right). \tag{93}$$

The preceding exponential estimate is weak when m is small. To obtain a constant lower probability uniformly in m, note that the characteristic function of Z_1, a difference of two independent $\mathrm{Bin}(N, q)$ variables, is $[1 - 2q(1-q)(1-\cos t)]^N$. Fourier inversion therefore gives, for every integer a,

$$\Pr(Z_1 = a) \le \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(1 - 2q(1-q)(1-\cos t)\right)^N \, dt.$$

The expression in parentheses is nonnegative. Using $1 - u \le e^{-u}$ and $1 - \cos t \ge 2t^2/\pi^2$ for $|t| \le \pi$ yields

$$\begin{aligned} \Pr(Z_1 = a) &\le \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\left(-\frac{2\tau^2 t^2}{\pi^2}\right) dt \\ &\le \frac{\sqrt{\pi}}{2\sqrt{2}\,\tau} \le \frac{1}{\tau}. \end{aligned}$$

For every real $u > 0$, the open interval $(-u, u)$ contains at most $2u + 1$ integers. Thus

$$\Pr(|Z_1| < u) \le \frac{2u+1}{\tau}.$$

Assume first that $m \le 256$ and set $u = k\sqrt{mN}$. Then

$$\frac{u}{\tau} \le \frac{4km}{\sqrt{7}} \le \frac{1}{16\sqrt{7}} < \frac{1}{40}, \qquad \frac{1}{\tau} < \frac{1}{20}.$$

Hence $\Pr(|Z_1| < u) < 1/10$. By symmetry,

$$\Pr(Z_1 \le -u) = \frac{1 - \Pr(|Z_1| < u)}{2} > \frac{9}{20}.$$

This identity holds for real u as well as integer u, so no rounding of the threshold is needed. Since $T \ge (-Z_1)_+/N$, we have proved

$$\Pr\left(T \ge k\sqrt{\frac{m}{N}}\right) > \frac{9}{20} \qquad (m \le 256).$$

For $m > 256$, (93) gives a lower probability at least $1 - e^{-7/4} > 2/5$. Therefore, for every $m \ge 2$,

$$\Pr\left(T \ge k\sqrt{\frac{m}{N}}\right) \ge \frac{2}{5}. \tag{94}$$

Each anchor count is binomial with mean $3N/4$. Hoeffding's inequality and a union bound give

$$\Pr\left(\min\{\widehat{p}_A(z), \widehat{p}_B(z)\} < \tfrac{1}{2}\right) \le 2e^{-N/8}. \tag{95}$$

Combining (94) and (95), and using $N \ge 2048$, proves $\Pr(\mathcal{E}) \ge 1/3$.

The exponential probability in the statement follows from the same event. If $m \le 256$, then $1 - e^{-m/1024} \le 1 - e^{-1/4} < 1/3$, so the constant lower bound already suffices. If $m > 256$, equations (93) and (95), together with $N \ge 1024m$, imply

$$\Pr(\mathcal{E}^c) \le e^{-7m/1024} + 2e^{-128m} \le e^{-m/1024}.$$

Dividing the last inequality by $e^{-m/1024}$ reduces it to

$$e^{-6m/1024} + 2e^{-(128 - 1/1024)m} \le 1,$$

which holds for $m > 256$. This proves (91). □

Theorem H.6. *Set $k = 2^{-14}$ and $c_* = k/24$. For every integer $m \ge 2$, integer $N \ge 1024m$, $\gamma \in [1/2, 1)$, and $\sigma \in [0, 1/2]$, there is a two-player common-interest robust Markov game with two actions per player, $S = m + 2$ states, $d_{\max} = m + 1$, rewards in $[0, 1]$, and $H_{\mathrm{R}} \le 2$, with the following property. Every primitive row, for both players, has the full-simplex ambiguity set*

$$\mathcal{P}_{i,s,\boldsymbol{a}} = \{q \in \Delta(\mathcal{S}) : \mathrm{TV}(q, P^0_{s,\boldsymbol{a}}) \le \sigma\}.$$

Draw N independent nominal observations per primitive row, independently across rows, including rows with identical population centers. There is a deterministic measurable exact empirical discounted robust Nash oracle whose output satisfies

$$\mathbb{P}\Big(\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \geq c_* \sqrt{m/N}\Big) \geq \max\{1/3,\, 1 - e^{-m/1024}\}. \tag{96}$$

The constants are independent of σ and γ in their stated ranges. Thus the conclusion includes the case where every uncertainty radius is the same strictly positive number. The construction can also be padded to any larger state space and any action cardinalities $A_1, A_2 \geq 2$, keeping the same conclusion and $d_{\max} = m + 1$.

Proof. **1. Specify the population game.** The state space is $\{x, z, 1, \ldots, m\}$. The state x is the hub, z is an anchor branch, and $1, \ldots, m$ are the active branches. Each player has actions $0, 1$; at x, player 1's actions are also written A, B. At every branch $j \in \{z, 1, \ldots, m\}$, set

$$r_j(0,0) = \frac{1}{2}, \qquad r_j(1,1) = 1, \qquad r_j(0,1) = r_j(1,0) = 0.$$

The nominal transition from every branch and every joint action is δ_x. At the hub, all four nominal rows equal p, where

$$p(z) = \frac{3}{4}, \qquad p(j) = \frac{1}{4m} \quad (1 \leq j \leq m), \qquad p(x) = 0.$$

All rewards are common to both players. Use $k = 2^{-14}$ from Lemma H.5, and put

$$c_N = \frac{\gamma k}{4}\sqrt{m/N}, \qquad r_x(A,0) = \frac{1}{2}, \qquad r_x(B,0) = \frac{1}{2} + c_N, \qquad r_x(A,1) = r_x(B,1) = 0.$$

Since $k \leq 1$ and $N \geq 1024m$, we have $0 < c_N \leq 1/4$. Every primitive row uses the same TV radius σ over the entire simplex $\Delta(\mathcal{S})$. In particular, nature can move probability from any branch to any other state; the argument does not restrict nature to the nominal support.

2. Define and verify an exact oracle on every sample realization. Write $\widehat{p}_{ab}$ for the independent empirical hub rows and $\Delta = \widehat{p}_{A0} - \widehat{p}_{B0}$. Define

$$b_z = 1, \qquad b_j = \mathbf{1}\{\Delta(j) \geq 0\} \quad (1 \leq j \leq m), \qquad \mathcal{H} = \{j \in \{z, 1, \ldots, m\} : b_j = 1\}.$$

Both players choose b_j at branch j. Let $u_{ab} = \widehat{p}_{ab}(\mathcal{H})$. At the hub, player 2 chooses 0, while player 1 chooses A if

$$-c_N + \frac{\gamma}{2}\big[(u_{A0} - \sigma)_+ - (u_{B0} - \sigma)_+\big] \geq 0, \tag{97}$$

and B otherwise. Ties are resolved in favor of A. The rule is Borel measurable because it uses only finitely many arithmetic operations and sign tests. To define it on the entire space of empirical centers, use the displayed rule when every branch center equals δ_x and every hub center is supported on the displayed branches. Outside this closed set, regard joint actions as the actions of one robust controller and choose the lexicographically first maximizing joint action at each state of its discounted optimality equation. The resulting deterministic joint policy is a product Nash equilibrium: a unilateral response is among the centralized controller's feasible policies. The centralized value is continuous in the centers by the standard discounted contraction and continuity of TV support functionals. Its finitely many action scores are therefore continuous, so lexicographic maximization is Borel measurable. This supplies a measurable exact oracle once the displayed rule is verified on its stated domain.

We verify optimality for *all* empirical samples, including samples whose anchor count is small. Let $\widehat{V}$ be the common robust discounted value of the selected profile, and write

$$U = \gamma \phi_{\delta_x, \sigma}(\widehat{V}), \qquad L = U + \frac{1}{2}, \qquad v_{\min} = \min_s \widehat{V}(s).$$

The empirical branch centers equal δ_x almost surely. Their radius is the same for every action, so policy evaluation gives

$$\widehat{V}(j) = L + \frac{1}{2} b_j \qquad (j = z, 1, \ldots, m).$$

For any probability center q supported on the branches, removal of the largest σ units of branch probability and placement of that mass at a global minimum gives

$$\phi_{q,\sigma}(\widehat{V}) = \sigma v_{\min} + (1-\sigma)L + \frac{1}{2}(q(\mathcal{H}) - \sigma)_+. \tag{98}$$

To obtain this formula from Lemma F.28, retain $1-\sigma$ units of nominal mass, taking low branches before high branches. The retained high mass is $(q(\mathcal{H}) - \sigma)_+$, so the retained expectation equals $(1-\sigma)L + (q(\mathcal{H}) - \sigma)_+/2$. The remaining σ units contribute $\sigma v_{\min}$. This calculation also covers the case with no low branches: then $q(\mathcal{H}) = 1$. The minimizing state can lie outside the nominal support, so the formula uses the full-simplex ambiguity set.

The first two terms of (98) do not depend on q. Consequently, player 1's empirical Bellman advantage of A over B against player 2's action 0 is exactly the left side of (97). The chosen action is a best response. If player 1 chooses $a \in \{A, B\}$, player 2's advantage of 0 over 1 is at least

$$\frac{1}{2} - \frac{\gamma}{2}(1-\sigma) > 0,$$

since $(u-\sigma)_+ \in [0, 1-\sigma]$ for $u \in [0,1]$. Thus the two independently estimated rows relevant to player 2's incentive are fully accounted for. At every branch, a unilateral mismatch changes reward $1/2$ or 1 to zero while leaving its entire ambiguity set unchanged. The selected action is strictly optimal there. Each player's policy value therefore satisfies its discounted optimal-response Bellman equation against the selected opponents. Proposition C.4 identifies this Bellman fixed point with the optimal response value. Thus the oracle is an exact discounted robust Nash equilibrium from every initial state, including against history-dependent unilateral deviations.

3. Force the population-inferior hub action with positive probability. Set

$$T = \sum_{j=1}^{m} (-\Delta(j))_+.$$

The low branches are precisely the active coordinates where $\Delta(j) < 0$. Since the two empirical rows have total mass one,

$$u_{A0} - u_{B0} = -\sum_{j:b_j=0} \Delta(j) = T.$$

On the event

$$\mathcal{E} = \left\{ \widehat{p}_{A0}(z) \geq \sigma, \widehat{p}_{B0}(z) \geq \sigma, T \geq k\sqrt{m/N} \right\}, \tag{99}$$

both u_{A0}, u_{B0} are at least σ, because $z \in \mathcal{H}$. The positive-part difference in (97) therefore equals T. Its left side is at least $-c_N + \gamma k\sqrt{m/N}/2 = c_N > 0$, so the oracle chooses A. Lemma H.5 gives $\mathbb{P}(\mathcal{E}) \geq \max\{1/3, 1 - e^{-m/1024}\}$: its anchor condition is $\widehat{p}(z) \geq 1/2 \geq \sigma$. Notice that the anchor creates a supply of removable high mass, while the empirical difference of its own count cancels through $\sum_s \Delta(s) = 0$.

4. Convert the hub error to a robust average-reward gap. On $\mathcal{E}$, compare the selected population profile π^A with the unilateral deviation π^B that changes only player 1's hub action to B. All population centers and radii are action independent at each state. Therefore the set of possible induced stationary state transition kernels is exactly the same for these two deterministic profiles. Denote this common set by $\mathcal{Q}$. For each $Q \in \mathcal{Q}$, the two state reward functions differ by $c_N \mathbf{1}_{\{x\}}$.

Put $\alpha = 1 - \sigma$. Every branch row of Q assigns mass at least α to x, because its center is δ_x. If $z_t = \mathbb{P}_Q(S_t = x)$, then $z_{t+1} \geq \alpha(1 - z_t)$: whenever the current state is a branch, the next state is x with probability at least α. Summing over a horizon n and using $\sum_{t=0}^{n-1} z_{t+1} = \sum_{t=0}^{n-1} z_t + z_n - z_0$ gives

$$(1+\alpha)\sum_{t=0}^{n-1} z_t \geq \alpha n + z_0 - z_n \geq \alpha n - 1.$$

Here n is a time horizon, distinct from the sample-noise statistic T in Step 3. Thus every initial state and every $Q \in \mathcal{Q}$ have limiting expected average hub frequency at least $\alpha/(1+\alpha) = (1-\sigma)/(2-\sigma)$.

The expected Cesàro averages converge for every finite stationary Markov chain, including periodic and multichain kernels. Hence

$$g_Q^{\pi^B}(s) \geq g_Q^{\pi^A}(s) + \frac{1-\sigma}{2-\sigma} c_N.$$

Taking the infimum over the *same* kernel class $\mathcal{Q}$ on both sides gives

$$\underline{g}^{\pi^B}(s) - \underline{g}^{\pi^A}(s) \geq \frac{1-\sigma}{2-\sigma} c_N \geq \frac{c_N}{3}.$$

Since π^B is an admissible unilateral deviation,

$$\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \geq \frac{\gamma k}{12}\sqrt{m/N} \geq \frac{k}{24}\sqrt{m/N} \qquad \text{on } \mathcal{E}.$$

The same argument applies to history-dependent nature. Replacing the prescribed hub action in each history identifies the admissible state-row processes under the two deterministic profiles. For any such common process, its finite-horizon reward difference is $c_N n^{-1} \sum_{t=0}^{n-1} z_t \geq c_N \alpha/(1+\alpha) - c_N/[n(1+\alpha)]$. Taking lower limits and then infima over this common process class gives the same lower bound.

5. Verify the span bound for every stationary opponent. Fix either player and an arbitrary stationary opponent policy. Its own action cannot affect the population ambiguity set at a state, so the response operator is

$$(Th)(s) = g_s + \phi_s(h), \qquad g_s = \max_{a_i} \sum_{a_{-i}} \theta(a_{-i} \mid s) r(s, a_i, a_{-i}) \in [0,1].$$

Here $\phi_s = \phi_{\delta_x,\sigma}$ at every branch and $\phi_x = \phi_{p,\sigma}$. For $v \in [0,2]$, define a bias candidate by $h_v(j) = g_j$ at each branch and $h_v(x) = v$, and define

$$F(v) = v + \phi_{\delta_x,\sigma}(h_v) - \phi_{p,\sigma}(h_v).$$

The support functionals are continuous, hence F is continuous. At $v = 0$, all coordinates of h_0 are nonnegative and $h_0(x) = 0$, so $\phi_{\delta_x,\sigma}(h_0) = 0$ and $F(0) \leq 0$. At $v = 2$, every row in the ball about δ_x retains probability at least $1 - \sigma$ at x, and the remaining bias coordinates are nonnegative. Thus

$$F(2) \geq 2 + 2(1-\sigma) - p^\top h_2 \geq 2,$$

because $p^\top h_2 \leq 1$ and $\sigma \leq 1/2$. The intermediate value theorem supplies $v_* \in [0,2]$ with $F(v_*) = g_x$. Set $h = h_{v_*}$ and $\rho = \phi_{\delta_x,\sigma}(h)$. At each branch, $\rho + h(j) = g_j + \phi_{\delta_x,\sigma}(h)$ by construction. At the hub, $F(v_*) = g_x$ gives $\rho + h(x) = g_x + \phi_{p,\sigma}(h)$. Thus (ρ, h) solves the response constant-gain Bellman equation, and $h \in [0,2]^{m+2}$ gives $\mathrm{sp}(h) \leq 2$. The construction is uniform over both players and every stationary opponent policy, proving $H_{\mathrm{R}} \leq 2$.

6. Pad actions and states without changing the lower bound. For $A_1, A_2 \geq 2$, assign common reward zero to every joint action containing an added action, and retain the same nominal centers and radius σ. Estimate every added hub row independently. Against the selected profile, any added branch action has zero reward and the same robust continuation, so it is strictly worse. An added hub action has zero reward and can improve the robust continuation by at most $(1-\sigma)/2$, whereas the selected action has reward at least $1/2$. It is therefore strictly worse as well. The population span proof uses only $g_s \in [0,1]$ and action-independent ambiguity sets, so it remains valid. To enlarge the state space, add branch states with the same coordination rewards and nominal row δ_x, and select $(1,1)$ there. The hub center assigns zero probability to these states. All empirical hub rows also assign them zero probability, so the oracle and noise calculations are unchanged. Nature may nevertheless enter these added states, since the TV balls remain full-simplex balls. The occupancy and span arguments already cover every branch state and therefore continue to apply. □

Lemma H.7. *Let $N_0 \geq 1$ be an integer, $S \geq 2$, and $0 \leq \sigma \leq 1/2$. There are two S-state, two-player, two-action common-interest games, with every primitive TV radius equal to σ, nominal row supports of size at most two, and $H_{\mathrm{R}} \leq 1$, such that the following holds. If $1 \leq N \leq N_0$ samples are drawn from every primitive row and $0 < \varepsilon \leq 1/(400\sqrt{N_0})$, every learner returning a stationary product profile has probability at least $3/8$ of returning a profile with robust average-reward Nash gap larger than ε in at least one of the two games.*

Proof. Put $a = 1/4$ and $u = 1/(100\sqrt{N_0})$. There is one hub x and $S-1$ branch states, of which z is distinguished. Both players receive reward zero at x and reward one at every branch, irrespective of their actions. Every nominal branch row is δ_x. Player 1 labels its hub actions A, B, and player 2 has two behaviorally identical actions. For a sign $\vartheta \in \{-1, 1\}$, define the nominal hub rows by

$$P^0_\vartheta(z \mid x, A, b) = \sigma + a + \vartheta u, \qquad P^0_\vartheta(z \mid x, B, b) = \sigma + a - \vartheta u, \qquad b \in \{0, 1\},$$

and place the remaining probability on x. Every primitive ambiguity set is the full-simplex TV ball of radius σ about its displayed center. These are valid probability rows, because $u \le 1/100$ and $\sigma + a \in [1/4, 3/4]$.

We first compute the robust gains, including for mixed profiles. Suppose player 1 plays A with probability w at the hub. Write

$$t_\vartheta(w) = a + \vartheta u(2w - 1), \qquad \rho_\vartheta(w) = \frac{t_\vartheta(w)}{1 + t_\vartheta(w)},$$

$$h(x) = 0, \qquad h(j) = \frac{1}{1 + t_\vartheta(w)} \quad (j \neq x).$$

For every branch row, the robust expectation of h is zero, since δ_x is already supported at its minimum. At an A hub row, the robust expectation is $(a + \vartheta u)h(z)$, because nature can remove exactly σ probability from z and place it at x. At a B hub row, the same calculation gives $(a - \vartheta u)h(z)$. Thus the robust policy-evaluation Bellman equations are

$$\rho_\vartheta(w) = t_\vartheta(w)h(z), \qquad \rho_\vartheta(w) + h(j) = 1 \quad (j \neq x).$$

The fixed-policy version of Proposition C.6 identifies $\rho_\vartheta(w)$ as the robust average payoff from every state. The calculation is unaffected by player 2's policy. Player 1's optimal response uses A when $\vartheta = 1$ and B when $\vartheta = -1$. The corresponding bias has span at most one. For player 2, every action has the same population reward and ambiguity set, so the displayed fixed-policy equation is also its optimal-response equation against any player 1 policy. Consequently $H_{\mathrm{R}} \le 1$.

If player 1 puts at most one-half of its probability on the correct hub action, its robust Nash gap is at least

$$\frac{a + u}{1 + a + u} - \frac{a}{1 + a} = \frac{u}{(1 + a + u)(1 + a)} > \frac{u}{2} > \varepsilon.$$

An ε-equilibrium therefore identifies the sign uniquely by which hub action receives probability greater than one-half.

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the joint laws of all sampled rows in the two games. Only the four hub rows differ. Use the elementary Bernoulli divergence bound from Step 4 of Theorem H.2. The parameters here differ by $2u$ and lie in $[0.24, 0.76]$. In particular, each directed row divergence is at most $64u^2$. Independence and additivity of relative entropy yield

$$\mathrm{KL}(\mathbb{P}_+ \| \mathbb{P}_-) \le 256 N u^2 \le \frac{256}{10000} < \frac{1}{8}.$$

Pinsker's inequality consequently gives $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) < 1/4$. The same bound remains true after adding independent internal randomness to the learner's input.

For any binary decision rule D,

$$\mathbb{P}_+(D = -) + \mathbb{P}_-(D = +) = 1 - \big[\mathbb{P}_+(D = +) - \mathbb{P}_-(D = +)\big] \ge 1 - \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) > \frac{3}{4}.$$

Decode the learner's output by its majority hub action, breaking a tie arbitrarily. Whenever the returned profile is an ε-equilibrium, the decoded sign is correct. Thus the sum of the two equilibrium-failure probabilities exceeds $3/4$, and one of them is at least $3/8$. □

Corollary H.8. *There are universal constants $c, \varepsilon_0 > 0$ such that the following holds. Let $S \ge 4$, $3 \le D \le S - 1$, $\sigma \in [0, 1/2]$, and $\gamma \in [1/2, 1)$. Consider the class of S-state, two-player, two-action common-interest games with $H_{\mathrm{R}} \le 2$, nominal support sizes at most D, and every primitive TV radius equal to σ. If a balanced empirical plug-in guarantee succeeds with probability at least $3/4$ for every game in this class and every measurable exact discounted robust Nash oracle, then, for $0 < \varepsilon \le \varepsilon_0$,*

$$N \ge c \frac{D}{\varepsilon^2}.$$

This necessity holds for every integer sample budget $N \geq 1$. Consequently, since $A_\times = 4$ here, the balanced protocol requires

$$N_{\mathrm{tot}} = SA_\times N = \Omega\left(SA_\times \frac{D}{\varepsilon^2}\right).$$

In particular $D = S - 1$ gives $\Omega(S^2 A_\times/\varepsilon^2)$, uniformly over all common radii $\sigma \in [0, 1/2]$, at constant span.

Proof. Put $m = D - 1 \geq 2$ and use $c_* = 2^{-14}/24$ from Theorem H.6. Choose

$$\varepsilon_0 = \min\left\{\frac{1}{400\sqrt{2048}}, \frac{c_*}{2\sqrt{2048}}\right\}.$$

The numerical constants are deliberately conservative; their size has no role in the asymptotic conclusion.

If $N \leq 2048$, Lemma H.7, with $N_0 = 2048$, contradicts the proposed guarantee. Its games have support at most two, so they belong to the class with support budget D. They also have $H_{\mathrm{R}} \leq 1 \leq 2$ and the required common radius.

Suppose next that $2048 < N < 1024m$, and set $m' = \lfloor N/1024 \rfloor$. Then $2 \leq m' \leq m$ and $N \geq 1024m'$. Since $N/1024 \geq 2$, the inequality $\lfloor t \rfloor \geq t/2$ for $t \geq 2$ gives $m'/N \geq 1/2048$. Apply Theorem H.6 with m' active branches, and pad with high branches to obtain exactly S states. Its nominal support is $m' + 1 \leq D$, and the theorem's padding argument preserves all ambiguity sets, span bounds, and incentives. With probability at least $1/3$, the selected profile has gap at least

$$c_* \sqrt{m'/N} \geq \frac{c_*}{\sqrt{2048}} \geq 2\varepsilon_0 > \varepsilon.$$

This again contradicts success probability at least $3/4$.

It remains that $N \geq 1024m$. Use the construction with m active branches and pad to S states. If $N < c_*^2 m/\varepsilon^2$, its gap exceeds ε with probability at least $1/3$, a contradiction. Thus $N \geq c_*^2 m/\varepsilon^2$. Since $D \geq 3$ gives $m = D - 1 \geq 2D/3$, the stated bound follows with $c = 2c_*^2/3$. Multiplication by the actual number $SA_\times$ of sampled primitive rows gives the total requirement. This multiplication comes from the balanced protocol; it does not assert that every row is informative. □

Remark H.9. *The dense construction also has a fixed exact equilibrium for every empirical sample: choose $(1, 1)$ at every branch, and $(B, 0)$ at the hub. Indeed, all branch values then equal the same number K, while every empirical hub center is supported on branches. If $v_* = \min_s V(s)$,*

$$\phi_{q,\sigma}(V) = (1 - \sigma)K + \sigma v_*$$

for every such center q. Hence all hub robust continuations coincide, and $(B, 0)$ is optimal because of its immediate rewards. Branch unilateral mismatches reduce the immediate reward and keep the same ambiguity set. The argument applies to the population centers as well, so this fixed profile is also a population discounted robust equilibrium. It is also an average-reward robust equilibrium. To verify this directly, put $b = 1/2 + c_N < 1$ and define

$$h(x) = 0, \qquad h(j) = \frac{1 - b}{2 - \sigma} \quad (j \neq x), \qquad \rho_\sigma = \frac{b + 1 - \sigma}{2 - \sigma}.$$

The bias is positive on branches, so $\phi_{\delta_x,\sigma}(h) = 0$ and $\phi_{p,\sigma}(h) = (1 - \sigma)h(j)$. At each branch $\rho_\sigma + h(j) = 1$, and at the hub $\rho_\sigma = b + (1 - \sigma)h(j)$. Since the chosen actions maximize each player's immediate reward against the other's fixed action, and population ambiguity sets are action independent, these equations are the optimal-response Bellman equations for both players. Their verification proves exact average-reward equilibrium. Moreover, holding this game's rewards fixed,

$$\frac{d\rho_\sigma}{d\sigma} = \frac{-(2 - \sigma) + (b + 1 - \sigma)}{(2 - \sigma)^2} = \frac{b - 1}{(2 - \sigma)^2} < 0.$$

Thus the positive ambiguity is payoff relevant; it strictly decreases even the stable equilibrium's robust average reward.

The lower bound therefore concerns a guarantee required to hold for every *exact empirical oracle. It does not give an S^2 minimax lower bound over all learners or all designed oracles. It proves that*

the successor factor survives a common strictly positive ambiguity radius even on the statistically difficult rows. This construction fixes the span while allowing a common radius on every row. The joint support–span–ambiguity dependence for the larger class with radii bounded above by σ is established separately in Corollary H.14 below.

H.2.2 Joint support–span–ambiguity lower bound and all-budget completion

We strengthen the successor lower bound by realizing its interaction with the response span and the ambiguity radius in one game. The sampling interface is unchanged: every primitive row is estimated independently from N nominal observations, and the guarantee is required to hold for every exact empirical discounted Nash oracle. The construction uses two players with two actions and common rewards. Its TV radii are at most σ; the nonzero radii are on the statistically difficult hub rows. The constant-common-radius result of Theorem H.6 remains a separate specialization.

The proof has three ingredients. Paired branch counts produce a sampling fluctuation of order $\sqrt{bm/N}$. An empirical equilibrium assigns the larger continuation value to the favorable member of each pair, amplifying that fluctuation by order $1/q$. Finally, a two-point construction and a reduction in the number of active pairs cover sample budgets below the range of the main construction. Since b/q^2 is of order $H + \sigma H^2$, these ingredients give the joint dependence.

Parameters. Throughout the large-span construction, set

$$\begin{aligned} &H_0 = 128, \qquad m_0 = 4096, \qquad C_0 = 2^{24}, \qquad \kappa = 2^{-12}, \\ &q = \frac{4}{H}, \quad r = \frac{\sigma}{4}, \quad b = q + r, \quad L_0 = \frac{b}{q^2} = \frac{H}{4} + \frac{\sigma H^2}{64}. \end{aligned} \tag{100}$$

For the large-span construction, m, N are integers satisfying $m \geq m_0$ and $N \geq C_0 m L_0$. We also assume $H \geq H_0$, $0 \leq \sigma \leq 1/2$, and $1 - 1/H \leq \gamma < 1$. Thus $0 < q \leq 1/32$, $0 \leq r \leq 1/8$, $q \leq b \leq 5/32$, and $\lambda = 1 - \gamma \leq q/4$. Define

$$u = \sqrt{\frac{bm}{N}}, \qquad c_N = \kappa \frac{u}{q} = \kappa \sqrt{\frac{mL_0}{N}}. \tag{101}$$

In particular, $u/q \leq C_0^{-1/2}$ and $0 < c_N \leq 2^{-24} < 1/4$. Throughout this subsection, write $\phi_{p,r}(v) = \min_{\mathrm{TV}(z,p) \leq r} z^\top v$, where the minimum is over the full state simplex.

Lemma H.10. *Let p be the distribution on $\{x\} \cup \{j^+, j^- : 1 \leq j \leq m\}$ with $p(x) = 1 - 2b$ and $p(j^+) = p(j^-) = b/m$. Let $\widehat{p}_{00}, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11}$ be independent empirical distributions from N observations each. Independently, let $\widehat{q}_{ab}$, $a, b \in \{0, 1\}$, be empirical means of N Bernoulli(q) observations. Write*

$$\Delta = \widehat{p}_{00} - \widehat{p}_{10}, \quad W = \sum_{j=1}^{m} |\Delta(j^+) - \Delta(j^-)|, \quad \Delta_B = \sum_{j=1}^{m} \{\Delta(j^+) + \Delta(j^-)\}.$$

Under the parameter conditions above, the event

$$\begin{aligned} \mathcal{E} = \Bigg\{ &W \geq u/2, \quad |\Delta_B| \leq u/8, \quad \max_{a,b} \sum_{j,\pm} |\widehat{p}_{ab}(j^\pm) - p(j^\pm)| \leq q/64, \\ &q/2 \leq \widehat{q}_{ab} \leq 2q \quad \textit{for all } a, b \Bigg\} \end{aligned} \tag{102}$$

has probability at least $3/4$.

Proof. We first lower-bound the mean of W. For a fixed pair j, put $Z_j = N\{\Delta(j^+) - \Delta(j^-)\}$. It is a sum of $2N$ independent, mean-zero variables taking values in $\{-1, 0, 1\}$, each with variance $2b/m$. Its variance is

$$v_j = \mathbb{E} Z_j^2 = 4Nb/m \geq 4C_0 b^2/q^2 \geq 4C_0 \geq 1.$$

The fourth-moment expansion, using independence and zero means, gives

$$\mathbb{E} Z_j^4 = \sum_k \mathbb{E} U_k^4 + 6 \sum_{k<\ell} \mathbb{E} U_k^2 \mathbb{E} U_\ell^2 \leq v_j + 3v_j^2 \leq 4v_j^2.$$

Here $U_k^4 \le U_k^2$, because $|U_k| \le 1$. Hölder's inequality gives $\mathbb{E}Z_j^2 \le (\mathbb{E}|Z_j|)^{2/3}(\mathbb{E}Z_j^4)^{1/3}$, and hence $\mathbb{E}|Z_j| \ge \sqrt{v_j}/2$. Summing expectations, without assuming independence between the coordinate counts, yields

$$\mathbb{E}W = \frac{1}{N}\sum_j \mathbb{E}|Z_j| \ge \frac{m}{N}\sqrt{Nb/m} = u.$$

We next control W about its mean. Replacing one of the $2N$ observations changes W by at most $2/N$. It changes nothing if both the old and new observations equal x. The probability that either observation is a branch is at most $4b$. Thus the mean squared replacement difference is at most $16b/N^2$. The independent-replacement Efron–Stein inequality (Boucheron et al., 2004) gives $\mathrm{Var}(W) \le \frac{1}{2}(2N)(16b/N^2) = 16b/N$. Chebyshev's inequality now implies

$$\mathbb{P}(W < u/2) \le \mathbb{P}(W < \mathbb{E}W/2) \le \frac{4\,\mathrm{Var}(W)}{(\mathbb{E}W)^2} \le \frac{64}{m}.$$

The total branch count in each empirical row is binomial with parameter $2b$. Therefore $\mathbb{E}\Delta_B = 0$ and $\mathrm{Var}(\Delta_B) = 4b(1-2b)/N \le 4b/N$, so

$$\mathbb{P}(|\Delta_B| > u/8) \le \frac{256}{m}.$$

For each hub row, Jensen's inequality and the binomial marginal variances give

$$\mathbb{E}\sum_{j,\pm} |\widehat{p}_{ab}(j^{\pm}) - p(j^{\pm})| \le 2m\sqrt{\frac{b}{mN}} = 2u.$$

Markov's inequality and a union bound over the four rows bound the failure probability of the third condition in (102) by $512u/q$. Finally, $\mathrm{Var}(\widehat{q}_{ab}) \le q/N$, and another union bound gives

$$\mathbb{P}\big(\exists a, b : \widehat{q}_{ab} \notin [q/2, 2q]\big) \le \frac{16}{Nq}.$$

The interval $[q/2, 3q/2]$ from the Chebyshev estimate is contained in the stated interval. Combining the bounds proves

$$\mathbb{P}(\mathcal{E}^c) \le \frac{320}{m} + \frac{512}{\sqrt{C_0}} + \frac{16}{C_0 m} \le \frac{320}{4096} + \frac{1}{8} + \frac{16}{2^{24}4096} < \frac{1}{4}.$$

For the penultimate inequality we used $Nq \ge C_0 mb/q \ge C_0 m$. □

Lemma H.11. *Consider a finite robust game with rewards in $[0,1]$. Suppose there is a state x such that, from every other state, the expected first hitting time of x is at most T under every joint-action strategy and every admissible nature strategy. Then every fixed-opponent response problem has a constant-gain solution, and $H_{\mathrm{R}} \le \max\{1, 2T\}$.*

Proof. Apply Corollary F.10 with the common anchor $z_{i,\theta} = x$. The hypothesis holds against every joint-action and nature strategy, so both anchor conditions hold with $D_{\mathrm{P},z} = D_{\mathrm{N},z} = T$; at x the hitting time is zero. Since rewards lie in $[0,1]$, $R_i \le 1$. The corollary supplies a constant-gain solution for every response problem and gives $H_{\mathrm{R}} \le \max\{1, 2T\}$. □

Theorem H.12. *Under the large-span parameter conditions above, there is a common-interest game with two players, two actions each, $2m+2$ states, nominal successor support $d_{\max} = 2m+1$, full-simplex rectangular TV radii at most σ, and*

$$H/16 \le H_{\mathrm{R}} \le H,$$

with a deterministic exact empirical discounted Nash oracle satisfying

$$\mathbb{P}\left(\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \ge \frac{\kappa}{3}\sqrt{\frac{mL_0}{N}}\right) \ge \frac{3}{4}. \tag{103}$$

The game may depend on N, and the oracle is exact on every possible sample realization. The same conclusion holds after padding to any larger state space. Since $L_0 \ge (H + \sigma H^2)/64$, the event also gives

$$\mathrm{Gap}_{\mathrm{av}}(\widehat{\pi}) \ge \frac{\kappa}{24}\sqrt{\frac{m(H+\sigma H^2)}{N}}.$$

Proof. Step 1: specify the complete game. The states are a hub x, a holding state y, and pairs j^+, j^-, $j \in [m]$. Both players have actions $0, 1$; player 1's hub actions are also called $A = 0$ and $B = 1$. Rewards are common to the two players. At x set

$$r_x(0,0) = 1/4, \quad r_x(1,0) = 1/4 + c_N, \quad r_x(0,1) = r_x(1,1) = 0.$$

All four nominal hub rows are the distribution p in Lemma H.10, with zero probability on y. Their TV radius is $r = \sigma/4$ for both players. At every branch the reward is zero. Joint action $(1,1)$ moves deterministically to y, and every other joint action moves deterministically to x. All branch radii are zero. At y every joint action has reward one, nominal row $q\delta_x + (1-q)\delta_y$, and radius zero. Padding states, if needed, have reward zero, radius zero, and transition δ_x under every action. All probabilities are valid because $2b \le 5/16 < 1$; all rewards lie in $[0,1]$. Only a hub row has support larger than two.

We first verify the span upper bound for every opponent policy. From y, the mean hitting time of x is $1/q$. From any branch it is at most $1 + 1/q$, regardless of the players' actions. Nature only acts at the hub, where the hitting time has already stopped. Lemma H.11 therefore gives

$$H_{\mathrm{R}} \le 2(1 + 1/q) = 2 + H/2 \le H.$$

This establishes the required uniformity over all stationary opponents, not just the opponents selected by the oracle.

Step 2: define an exact oracle on every sample realization. Let $\widehat{p}_{ab}$ be the empirical hub rows and $\widehat{q}_{ab}$ the four empirical exit probabilities at y. All deterministic rows are estimated exactly with probability one. For each pair, mark j^+ high if $\Delta(j^+) \ge \Delta(j^-)$ and mark j^- high otherwise. Let $\mathcal{H}$ be the set of high branches; the other member of each pair is low. At high branches play $(1,1)$ and at low or padding branches play $(0,0)$. At y choose the lexicographically first joint action attaining $\widehat{q} = \min_{a,b} \widehat{q}_{ab}$.

Fix these branch and holding-state choices, and fix player 2's hub action to zero. Solve the restricted discounted control problem in which only player 1's hub action remains free. Proposition C.4 supplies its value and a deterministic optimal hub action, with ties broken lexicographically. These choices define a candidate profile. Evaluate its common robust discounted value and test, at every state, that every unilateral action has Bellman score at most that value. If these inequalities hold, return the candidate. By Proposition C.4, this test certifies exact statewise discounted Nash optimality. Otherwise solve the centralized robust control problem whose action set is the set of four joint actions, and return a deterministic greedy joint policy, breaking ties lexicographically. This fallback is an exact empirical Nash equilibrium. Indeed, both players have identical rewards and identical ambiguity sets, and any unilateral strategy against the fixed opponent is also an admissible centralized joint strategy. The centralized optimum therefore upper-bounds the value of every unilateral deviation, while the selected joint policy attains that optimum. Existence of this centralized value and a stationary deterministic optimal joint policy is another application of Proposition C.4. A deterministic joint policy is a stationary product profile. Thus the fallback requires no separate general-sum equilibrium-existence assumption. For fixed N, the empirical kernels have finite range, so the prescribed choices are deterministic and measurable. The rule uses only the empirical game, the known rewards, and fixed state labels.

We show that on $\mathcal{E}$ the candidate passes the test and chooses A. Thus no properties of the fallback enter the probability calculation.

Step 3: analyze the candidate's value levels on $\mathcal{E}$. Write V for the restricted optimal value vector, and put $X = V(x)$ and $Y = V(y)$. The candidate, which is greedy for this restricted problem, has this same evaluation vector. Its low and high branch values are respectively

$$L = \gamma X, \qquad U = \gamma Y, \qquad Y = 1 + \gamma\{(1-\widehat{q})Y + \widehat{q}X\}.$$

Padding states, when present, also have value L. Bounded rewards give all values in $[0, 1/\lambda]$. The chosen hub reward R is in $[1/4, 1/2]$, so $X \ge R > 0$. If $\lambda X = 1$, the hub equation and $\phi_{\widehat{p},r}(V) \le 1/\lambda$ would give $1 \le \lambda R + \gamma$, or $1 \le R$, a contradiction. Consequently $0 < \lambda X < 1$. Put $c = \lambda X$. The holding-state equation gives

$$Y - X = \frac{1-c}{\lambda + \gamma\widehat{q}} > 0. \tag{104}$$

Indeed, subtracting X from the displayed holding equation for Y gives $(\lambda+\gamma\widehat{q})(Y-X) = 1-\lambda X = 1-c$. We next show $L < X < U < Y$. The inequalities $L < X$ and $U < Y$ follow from $L = \gamma X$, $U = \gamma Y$, $X > 0$, and $Y > 0$. Only $X < U$ requires an argument. If $U \le X$, all nominal hub successors would have value at most X, implying $X \le R + \gamma X$ and hence $c \le 1/2$. But $\gamma Y \le X$, substituted into the holding-state equation, implies

$$\gamma \le c(1+\gamma\widehat{q}), \qquad c \ge \frac{\gamma}{1+\gamma\widehat{q}} > \frac{1}{2},$$

because $\gamma \ge 1 - 1/128$ and $\widehat{q} \le 2q \le 1/16$. This is a contradiction. The global minimum is therefore a low branch, and the global maximum is Y.

Every selected high set has population mass b, because it contains one member of each equal-mass pair. Although this set depends on the samples, the coordinate ℓ_1 event bounds its mass error uniformly over all subsets of branches. Thus, on $\mathcal{E}$, each empirical hub row has high mass $h_{ab} = \widehat{p}_{ab}(\mathcal{H})$ satisfying $|h_{ab} - b| \le q/64$. In particular, $h_{ab} > r$. Let t_{ab} be its total branch mass. Nature removes r probability from high branches and puts it on a low branch. There is sufficient donor mass because $h_{ab} > r$, and sufficient receiver capacity because the total branch mass is at most $2b + q/64 < 1 - r$. It is optimal because high branches are the largest values in the nominal support and low branches are global minima. Lemma F.28 therefore gives the exact minimum directly. Hence

$$\phi_{\widehat{p}_{ab},r}(V) = (1 - t_{ab})X + (h_{ab} - r)U + (t_{ab} - h_{ab} + r)L. \tag{105}$$

For the chosen hub row put $a = h_{ab} - r$, $t = t_{ab}$, and $K = \gamma^2 a/(\lambda + \gamma\widehat{q})$. The hub equation and (105) yield

$$c(1+\gamma t) = R + \gamma^2 a(Y - X) = R + K(1 - c), \qquad c = \frac{R + K}{1 + \gamma t + K}.$$

For the first identity, substitute $L = \gamma X$, $U = \gamma Y$, and $\phi = (1-t)X + aU + (t-a)L$ into $X = R + \gamma\phi$, then subtract γX from both sides, use $c = \lambda X$, and apply (104). The second identity is the resulting linear equation solved for c. On $\mathcal{E}$ we have $0 < a \le 2q$ and $\widehat{q} \ge q/2$, so $0 < K \le 4$. Consequently $c \le (1/2 + K)/(1 + K) \le 9/10$. Equation (104) now implies, with $g = U - L = \gamma(Y - X)$,

$$\frac{1}{32q} \le g \le \frac{2}{q}, \qquad g \ge 1. \tag{106}$$

For the lower inequality use $1 - c \ge 1/10$, $\lambda + \gamma\widehat{q} \le 9q/4$, and $\gamma \ge 3/4$; their combination gives the stronger lower bound $1/(30q)$. For the upper inequality use $1 - c \le 1$ and $\lambda + \gamma\widehat{q} \ge \gamma q/2$. Finally $q \le 1/32$ proves $g \ge 1$.

Step 4: verify all empirical incentives. At each pair we selected the larger coordinate of Δ, so

$$\Delta(\mathcal{H}) = \frac{1}{2}(\Delta_B + W) \ge 3u/16, \qquad \Delta(x) = -\Delta_B, \quad |\Delta(x)| \le u/8.$$

The TV penalty in (105) is identical for both hub rows. Since $X - L = c$, player 1's Bellman advantage of A over B against action zero of player 2 is

$$\begin{aligned}
\mathcal{A}_{A-B} &= -c_N + \gamma\{\Delta(x)c + \Delta(\mathcal{H})g\} \\
&\ge -c_N + \gamma u(3g/16 - 1/8) \\
&\ge -c_N + \gamma g u/16 \\
&\ge \left(\frac{3}{2048} - \kappa\right)\frac{u}{q} > 0.
\end{aligned}$$

The second line uses $0 \le c \le 1$; the third uses $g \ge 1$; and the fourth uses (106) and $\gamma \ge 3/4$. Thus the restricted optimum selects A strictly.

For player 2, any two empirical hub rows are within $q/32$ in TV. Indeed, the total variation from each row to p is no larger than its branch coordinate ℓ_1 error: the hub-coordinate error is the negative of the sum of the branch errors. Use the triangle inequality and (102). After cancellation of the common TV penalty, the continuation difference between the two hub actions is therefore at most $(q/32)(U - L) \le 1/16$. The immediate reward advantage of action zero is at least $1/4$, so its Bellman

advantage is at least $1/4 - \gamma/16 > 0$. At a high branch, deviating changes the successor from y to x and strictly decreases value because $Y > X$. At a low branch, a unilateral deviation leaves the zero reward and successor x unchanged. At y, every unilateral alternative has exit probability at least $\widehat{q}$ and hence no larger continuation value because $Y > X$. Padding states are indifferent. The candidate's policy-evaluation vector consequently satisfies each player's full discounted optimal-response equation. Contraction proves exact Nash optimality, so the candidate passes the oracle's test on $\mathcal{E}$.

Step 5: evaluate the population average-reward deviation. Hold the selected high/low branch pattern fixed. Compare the population profiles with hub action $(A, 0)$ and $(B, 0)$, keeping the other actions fixed. For either hub reward $R \in \{1/4, 1/4 + c_N\}$ define

$$\rho(R) = \frac{1+R}{2+2b}, \qquad h(x) = 0, \quad h(y) = \frac{1-\rho(R)}{q}, \quad h(j) = \begin{cases} h(y) - \rho(R), & j \in \mathcal{H}, \\ -\rho(R), & j \notin \mathcal{H}. \end{cases} \tag{107}$$

Padding states also have bias $-\rho(R)$. Since $R \le 1/2$, $\rho(R) \le 3/4$, and $q \le 1/32$, $h(y) - \rho(R) > 0$. Thus low branches are global minima and high branches are the largest nominal hub successors. Their nominal mass is $b > r$. Nature optimally moves r high mass to low branches, giving

$$\phi_{p,r}(h) = (b-r)\{h(y) - \rho(R)\} + (b+r)\{-\rho(R)\} = qh(y) - 2b\rho(R).$$

The hub equation is $R + qh(y) - 2b\rho(R) = R + 1 - (1+2b)\rho(R) = \rho(R)$. The branch equations are $\rho + h(j) = h(y)$ at high branches and $\rho + h(j) = 0$ at low or padding branches. The holding equation is $\rho + h(y) = 1 + (1-q)h(y)$. These are fixed-policy robust constant-gain equations. Apply Proposition C.6 to the singleton-action robust MDP induced by each fixed profile. Its compact rectangular row sets attain the displayed minima, so its robust gain is $\rho(R)$ from every state under both nature conventions. Changing only player 1's hub action therefore improves gain by

$$\rho(1/4 + c_N) - \rho(1/4) = \frac{c_N}{2+2b} \ge \frac{c_N}{3},$$

where $2 + 2b \le 3$. On $\mathcal{E}$ the oracle chooses A, so this is a lower bound on its population Nash gap. Lemma H.10 proves (103). The same verification covers history-dependent nature.

Step 6: verify that the span parameter is genuinely of order H. Choose any fixed high/low pattern with one high branch per pair, and fix player 2 to action zero at the hub, action one at high branches, and action zero at low branches. Its action at y is immaterial in the population. This opponent policy is fixed independently of the samples. The witness (107) with hub action B is also an optimal-response witness for player 1: hub centers are action independent, B maximizes reward, high branches optimally lead to y, and low branches have identical unilateral transitions. Its gain is $\rho(1/4 + c_N) \le 3/4$. Every optimal-response Bellman solution has this same gain by Proposition C.6. At y, its equation necessarily gives $q\{h(y) - h(x)\} = 1 - \rho \ge 1/4$. Every such bias therefore has span at least $1/(4q) = H/16$. Taking the minimum over biases and the supremum over opponents proves the claimed lower bound on H_{R}. □

Lemma H.13. *Let $H \ge 128$, $0 \le \sigma \le 1/2$, and let $S \ge 2$ and $N \ge 1$ be integers. Define q, r, b, L_0 by* (100). *For $0 < e \le 1/64$ there are two S-state, two-player, two-action common-interest games, with radii at most σ, support at most two, and $H_{\mathrm{R}} \le H$, for which the following holds. If*

$$N \le c_T L_0 / e^2, \qquad c_T = 2^{-18}, \tag{108}$$

any learner using N independent observations of each primitive row and returning a stationary product profile has probability at least $3/8$ of returning a profile with average-reward Nash gap larger than e in one of the two games.

Proof. Put $\delta = 16eq \le q/4$ and let $\vartheta \in \{-1, 1\}$ be the hidden sign. The states are x, y and $S - 2$ padding states. At x both rewards are zero and the probability of moving to y is $b + \vartheta\delta$ for player 1's action A and $b - \vartheta\delta$ for its action B, independently of player 2's action. The remaining probability stays at x, and the hub radius is r. At y both rewards are one and every joint action has the zero-radius row $q\delta_x + (1-q)\delta_y$. All padding states have zero radius, deterministic transition to x, and common known reward $\rho_* = (q+\delta)/(2q+\delta)$, which is independent of the hidden sign.

For a mixed response placing probability w on A at x, set

$$p_w = q + \vartheta\delta(2w-1), \qquad \rho_w = \frac{p_w}{p_w+q},$$
$$h(x) = 0, \qquad h(y) = \frac{1}{p_w+q}, \qquad h(z) = \rho_* - \rho_w \quad \text{at padding states.}$$

Since $p_w \in [q-\delta, q+\delta]$ and the function $p/(p+q)$ is increasing for $p>0$, we have $p_w \geq 3q/4$ and $0 \leq h(z) = \rho_* - \rho_w \leq 1$. Also $p_w + q \leq 2q + \delta < 1$, so $h(y) > 1$. Thus x is a global bias minimum, and nature minimizes each hub expectation by moving exactly r probability from y to x. The two core Bellman equations are $\rho_w = p_w h(y)$ and $\rho_w + h(y) = 1 + (1-q)h(y)$; the padding equations follow from their chosen rewards. Verification gives the robust gain ρ_w from every state. Player 1's optimum uses the better hub action, and player 2's value is independent of its own actions. For player 1, against any opponent, use the witness for its better pure hub action. For player 2, against any stationary player-1 policy, use the witness for that policy's hub mixing weight w; player 2's own action is irrelevant. Their minimum bias is zero, and $h(y)$ is their maximum bias. Thus every response problem has an optimal bias of span at most $1/(2q-\delta) \leq 4/(7q) \leq H$. This proves the required opponent-uniform bound on H_{R}, including the defining floor at one.

If at most one-half of the mass is on the correct hub action, then $p_w \leq q$. The resulting gap is at least

$$\rho_* - \frac{1}{2} = \frac{\delta}{2(2q+\delta)} \geq \frac{\delta}{(9/2)q} > e.$$

An e-equilibrium must therefore put strictly more than one-half of its hub mass on the correct action. Decode $\vartheta = 1$ when $w > 1/2$ and $\vartheta = -1$ otherwise. Every decoding error implies a Nash gap larger than e, including the tie $w = 1/2$ under $\vartheta = 1$.

Only the four hub sample streams differ between the two games. Each has Bernoulli parameters $b+\delta$ and $b-\delta$. Both lie in $(0,1)$, and $v(1-v) \geq b/4$ for either parameter v, because $\delta \leq b/4$ and $b \leq 1/4$. The inequality $\log t \leq t-1$ gives

$$\mathrm{KL}(\mathrm{Ber}(u) \| \mathrm{Ber}(v)) \leq \frac{(u-v)^2}{v(1-v)} \leq \frac{16\delta^2}{b}.$$

Rewards and all other row laws are identical under the two signs, so they provide no additional information. Independence of the row samples bounds the KL divergence between the complete data laws by

$$64N\delta^2/b = 16384Ne^2/L_0 \leq 1/16.$$

Pinsker's inequality bounds the total variation by $\sqrt{1/32} < 1/4$. For every binary decision rule the sum of its two error probabilities is at least one minus this total variation, and hence exceeds $3/4$. Decoding the learner's majority hub action proves the claimed failure bound. Adding a common independent random seed for a randomized learner leaves the data-law KL divergence unchanged. The argument therefore covers randomized learners as well. □

Corollary H.14. *There are universal constants $c, \varepsilon_0 > 0$ such that the following holds for every $S \geq 4$, $3 \leq D \leq S-1$, $H \geq 1$, $0 \leq \sigma \leq 1/2$, and $0 < \varepsilon \leq \varepsilon_0$. Set $\gamma = 1 - \varepsilon/(2H)$. Here D is an integer support budget. Consider the class of two-player, two-action common-interest games with $H_{\mathrm{R}} \leq H$, nominal support sizes at most D, and rectangular full-simplex TV radii at most σ. If a balanced empirical-model guarantee returns an ε-average-reward Nash equilibrium with probability at least $3/4$ for every game and every exact measurable empirical discounted Nash oracle, then its per-row sample budget necessarily satisfies*

$$N \geq cD\frac{H+\sigma H^2}{\varepsilon^2}, \qquad N_{\mathrm{tot}} = 4SN = \Omega\left(SD\frac{H+\sigma H^2}{\varepsilon^2}\right). \tag{109}$$

Equivalently, for every integer budget below the displayed lower bound, there is a game and an exact empirical oracle whose failure probability exceeds $1/4$. The necessity thus holds for every integer $N \geq 1$.

Proof. We first accommodate every $H \geq 1$ by a bounded reward rescaling. Put $\bar{H} = \max\{H, 128\}$, $q = 4/\bar{H}$, $r = \sigma/4$, $b = q + r$, and $L_0 = b/q^2$. Multiply all rewards in the constructions above by

$$a = \min\{1, H/(2+\bar{H}/2)\}, \qquad a_{\min} = 1/66.$$

Then $a \geq a_{\min}$. Discounted and average policy values, Nash gaps, and bias vectors scale by a; exact discounted equilibrium sets are unchanged. Because H_{R} includes a floor at one, it does not in general scale exactly by a. Instead, scaling the witnesses in Lemma H.11 gives $H_{\mathrm{R}} \leq \max\{1, a(2+\bar{H}/2)\} \leq H$ for the scaled games. The two-point games satisfy the same upper bound. Directly,

$$a^2 L_0 \geq \frac{H + \sigma H^2}{256}. \tag{110}$$

If $a = 1$, use $\bar{H} \geq H$ and $L_0 = \bar{H}/4 + \sigma\bar{H}^2/64$. If $a < 1$, then $\bar{H} = 128$, $a = H/66$, and $a^2 L_0 = (32H^2 + 256\sigma H^2)/66^2$. Its first term is at least $H/256$ because $H \geq 1$ and $32 \cdot 256 > 66^2$; its second is at least $\sigma H^2/256$.

Let $c_g = \kappa/3$ and choose any positive ε_0 no larger than

$$\min\left\{\frac{a_{\min}}{64}, \quad a_{\min}\sqrt{\frac{c_T}{2C_0 m_0}}, \quad \frac{a_{\min} c_g}{2\sqrt{2C_0}}\right\}.$$

These conservative numerical constants are independent of every problem parameter. This choice ensures $\varepsilon/a \leq 1/64$ and $\gamma \geq 1 - 1/\bar{H}$: for $H < 128$, use $\varepsilon \leq 1/64 \leq 2H/128$, and for $H \geq 128$ use $\varepsilon \leq 1 < 2$. Set $m = \lfloor (D-1)/2 \rfloor$, so $m \geq D/4$ and $2m + 2 \leq S$.

Small support budgets. If $m < m_0$, the scaled two-point lemma implies the necessary inequality $N > c_T a^2 L_0/\varepsilon^2$: otherwise its failure probability is at least $3/8$, even for arbitrary learners. Since $D \leq 2m_0$, this already has the required order by (110). Suppose henceforth that $m \geq m_0$.

Budgets below the main construction's range. If $N < C_0 m_0 L_0$, our choice of ε_0 gives $N < c_T a^2 L_0/\varepsilon^2$, so the scaled two-point lemma contradicts the proposed guarantee. If $C_0 m_0 L_0 \leq N < C_0 m L_0$, set $m' = \lfloor N/(C_0 L_0) \rfloor$. Then $m_0 \leq m' < m$, $N \geq C_0 m' L_0$, and $m' L_0/N \geq 1/(2C_0)$. Apply Theorem H.12 with m' pairs and pad to S states. Its scaled gap is at least $a c_g/\sqrt{2C_0} \geq 2\varepsilon_0 > \varepsilon$ with probability at least $3/4$, another contradiction.

Budgets in the main construction's range. It remains that $N \geq C_0 m L_0$. The construction with m pairs implies that $N < a^2 c_g^2 m L_0/(4\varepsilon^2)$ produces gap larger than 2ε with probability at least $3/4$. Therefore $N \geq a^2 c_g^2 m L_0/(4\varepsilon^2)$ is necessary. Using $m \geq D/4$ and (110) completes all cases. For example one may take $c = \min\{c_T/(2m_0), c_g^2/16\}/256$. The total count is $4SN$ because the balanced experiment samples every primitive row, including the uninformative ones. □

I Statistical interpretation and oracle design

This section separates the learning guarantees by their oracle and sampling quantifiers, proves three constructive selection results, and explains the source of the successor-support factor. The lower-bound constructions are in Appendix H.

I.1 Scope and tightness of the two guarantees

The upper theorem, adaptive control lower bound, and oracle-uniform lower bounds use different quantifiers. The following statements keep their implications separate.

I.1.1 The quantifiers behind the two learning questions

Let $\mathcal{C}$ be a fixed class of games, and let Sel_γ denote measurable rules returning an exact empirical discounted Nash equilibrium. A rule may use the known rewards and radii as well as the empirical kernel. For a balanced sample size N, define

$$L_N(\mathcal{M}, \mathfrak{D}) := \mathbb{P}_{\mathcal{M}}\left\{\mathrm{Gap}_{\mathrm{av}}\left(\mathfrak{D}(\widehat{\mathcal{M}}_N)\right) > \varepsilon\right\}.$$

The every-oracle guarantee asks for

$$\sup_{\mathcal{M}\in\mathcal{C}} \sup_{\mathfrak{D}\in\mathsf{Sel}_\gamma} L_N(\mathcal{M}, \mathfrak{D}) \leq \delta.$$

Designing a plug-in learner instead asks whether

$$\inf_{\mathfrak{D}\in\mathsf{Sel}_\gamma} \sup_{\mathcal{M}\in\mathcal{C}} L_N(\mathcal{M}, \mathfrak{D}) \leq \delta.$$

The chosen rule in the second display must work throughout $\mathcal{C}$; it is not allowed to know the unknown population kernel. Theorem H.12 constructs a game and a bad oracle, establishing the first obstruction. It does not lower-bound the second quantity. An impossibility theorem for every designed oracle would instead need, for every proposed selection rule, a game on which that rule fails. An unrestricted minimax theorem must quantify over an even larger set of learners, including adaptive sampling and outputs that need not be exact equilibria of the raw empirical model.

The upper theorem proves a single event on which all empirical equilibria generalize. At fixed $\mathcal{M}$ and N, this simultaneous formulation and the worst-oracle formulation are equivalent: the empirical kernel has finite range, so one can choose a bad equilibrium at each empirical model for which one exists, and use any exact oracle elsewhere. This modification on a finite set remains measurable. Consequently,

$$\sup_{\mathfrak{O}\in\mathsf{Sel}_\gamma} L_N(\mathcal{M},\mathfrak{O}) = \mathbb{P}_{\mathcal{M}}\Big\{\exists \pi \in \mathrm{NE}_\gamma(\widehat{\mathcal{M}}_N) : \mathrm{Gap}_{\mathrm{av}}(\pi) > \varepsilon\Big\}.$$

Here NE_γ is the set of exact stationary empirical discounted equilibria. The equivalence concerns the adversarial supremum over rules; a designed learner must still choose one rule before the population model is known.

A constructive separation within the hard family. The distinction is not merely logical. The bad oracle uses dense-row noise to assign different continuation levels to different branches. Proposition I.3 fixes all branches at the low equilibrium and still returns an exact empirical discounted Nash equilibrium on every sample realization. Only the scalar total branch masses then enter its population incentive error. Their concentration gives $\mathbb{P}\{\mathrm{Gap}_{\mathrm{av}} > \varepsilon\} \le e^{-16N\varepsilon^2}$. This proposition is restricted to the explicit family in Theorem H.12; it does not solve selection for the entire class in Corollary H.14. Proposition I.4 provides the broader shared-ambiguity team result, with the additional unichain assumption retained in its statistical statement.

I.1.2 Unrestricted generative-model learning

Theorem H.2 supplies the lower scale

$$\Omega\left(SA_\times \frac{H+\sigma H^2}{\varepsilon^2}\right)$$

in its stated parameter regime. Its one-active-player embedding applies to the action-budget class, which allows the budget $A_\times$ to be allocated to one player. For a prescribed action-size vector $(A_1,\dots,A_n)$, that embedding gives $\max_i A_i$ in place of $A_\times$. Thus the latter comparison leaves both a successor-support gap and an action-dependence gap. The unrestricted dense general-sum minimax rate remains unresolved.

Corollary I.1. *Under the parameter restrictions of Theorems H.2 and G.11, on the action-budget class with $d_{\max} = O(1)$,*

$$N^\star_{\mathrm{tot}} = \widetilde{\Theta}\left(SA_\times \frac{H+\sigma H^2}{\varepsilon^2}\right).$$

Here $N^\star_{\mathrm{tot}}$ is the infimum, over learners succeeding with probability at least $3/4$ throughout the class, of their worst-case expected number of transition observations. The support cap is a fixed constant at least two. Theorem H.2 supplies the lower bound with two-successor rows, and Theorem G.11 supplies the upper bound with a deterministic sample budget.

I.1.3 Balanced plug-in learning uniform over exact oracles

For $S \ge 4$, $3 \le D \le S-1$, and the accuracy range in Corollary H.14, that corollary establishes $N = \Omega(D(H+\sigma H^2)/\varepsilon^2)$ on the class with successor support at most D, two players, and two actions per player. Combined with Theorem G.11, this gives the joint rate

$$N_{\mathrm{tot}} = \widetilde{\Theta}\left(SD \frac{H+\sigma H^2}{\varepsilon^2}\right)$$

for balanced empirical-model guarantees uniform over every exact empirical Nash oracle, at fixed numerical confidence and sufficiently small universal accuracy tolerance. This is a fixed-action

statement: $A_\times = 4$. Multiplication by $4S$ comes from the balanced protocol, which samples uninformative rows as well as informative rows.

The joint factor is proved in one game, not by multiplying the earlier control and successor lower bounds. The holding-state exit probability and hub TV radius jointly produce nominal mass $b = \sigma/4 + 4/H$, adverse mass $4/H$, and continuation separation of order H. Paired branch selection then converts these quantities into one population incentive gap. The construction allows zero-radius rows. A common strictly positive radius at every primitive row is a distinct subclass, in which $H_{\mathrm{R}} \le 1/\sigma$; its fixed-span successor obstruction remains Theorem H.6.

Proposition I.2 gives a sample-free population equilibrium, and Proposition I.3 proves a state-linear bound for a specified exact empirical discounted oracle on the same family. These are separate claims: a population equilibrium need not be an equilibrium of a sampled model. The latter proposition shows directly that the lower construction does not obstruct every designed discounted plug-in oracle.

I.2 Designed equilibrium oracles

These results explain the distinction between a lower bound for every-oracle guarantees and the performance of a specified selection rule.

I.2.1 A sample-free population equilibrium

Proposition I.2. *The family in Theorem H.12 has a fixed exact population average-reward Nash equilibrium: play $(B, 0)$ at the hub, $(0, 0)$ at every branch and padding state, and any fixed joint action at y. No transition observations are needed to output this profile on the constructed family.*

Proof. Write $R = 1/4 + c_N$. We first evaluate the proposed profile by setting

$$\rho_0 = \frac{R}{1 + 2b + r}, \qquad h(x) = 0, \qquad h(y) = \frac{1 - \rho_0}{q}, \qquad h(j) = -\rho_0$$

at every branch and padding state. The nominal hub has mass $2b$ on low branches and $1 - 2b > r$ at x. Nature moves r hub mass to a low branch, so its minimum bias expectation is $-(2b + r)\rho_0$. The hub equation is $\rho_0 = R - (2b + r)\rho_0$. At a branch or padding state it is $\rho_0 - \rho_0 = 0 + h(x)$, and at y it is $\rho_0 + h(y) = 1 + qh(x) + (1 - q)h(y)$. Against opponent action zero, either own branch action leads to x with reward zero. At the hub the chosen actions maximize each player's immediate reward against the other player's action, and the population ambiguity sets are action independent. At y all actions coincide. Thus the displayed fixed-policy equations are also both players' optimal-response equations. Bellman verification proves exact Nash optimality from every initial state. The same profile works after positive reward rescaling and state padding. □

I.2.2 An exact empirical oracle on the lower-bound family

Proposition I.3 (A designed discounted plug-in oracle on the lower-bound family)**.** *Fix any game with the transition structure of Theorem H.12, including its possible padding states, and replace the hub reward margin c_N by any fixed $\eta \in (0, 1/4]$. Thus the two positive hub rewards are $1/4$ and $1/4 + \eta$, all hub centers coincide, and every branch has zero reward and moves to x unless both players choose action one. For every $\gamma \in (0, 1)$, there is a deterministic measurable exact empirical discounted Nash oracle $\mathfrak{O}_\gamma^{\mathrm{low}}$ such that, for every integer $N \ge 1$, under N independent nominal observations per primitive row,*

$$\Pr\Big(\mathrm{Gap}_{\mathrm{av}}(\mathfrak{O}_\gamma^{\mathrm{low}}(\widehat{P}^0)) > \varepsilon\Big) \le \exp(-16N\varepsilon^2), \qquad \varepsilon > 0. \tag{111}$$

In particular, for $0 < \varepsilon \le 1$ and $0 < \delta < 1/2$, this family admits a designed discounted plug-in learner with balanced total sample count $O(S\varepsilon^{-2}\log(1/\delta))$, independently of the nominal successor support size. The assertion concerns this constructed family, not the full class in Corollary H.14.

Proof. Write $r = \sigma/4$, $q = 4/H$, and $b = q + r$ as in Theorem H.12. In particular, $q > 0$ and $2b + r < 1$. We use $A = 0$ and $B = 1$ for player 1's hub actions.

1. Specify the rule using empirical quantities. At every branch and padding state choose $(0, 0)$. Every branch row used by this profile or by a unilateral branch deviation has empirical center δ_x on every possible sample realization, and its ambiguity radius is zero. Padding rows have the same property. At the holding state y, choose a joint action minimizing its empirical exit probability $\widehat{q}_{ab}$, with a fixed tie rule, and write $\widehat{q} = \min_{a,b} \widehat{q}_{ab}$. For each hub joint action, define

$$t_{ab} = \widehat{P}^0(\{j^+, j^- : j \in [m]\} \mid x, a, b), \qquad \alpha_{ab} = \min\{t_{ab} + r, 1\}.$$

Choose a hub joint action maximizing

$$\frac{R_{ab}}{1 + \gamma\alpha_{ab}}, \qquad R_{00} = 1/4, \quad R_{10} = 1/4 + \eta, \quad R_{01} = R_{11} = 0, \tag{112}$$

again with a fixed tie rule. Its second component is always zero, since both ratios with second component one are zero and the other two are strictly positive. These finite arithmetic and minimization operations define a deterministic measurable rule.

2. Verify exact empirical optimality from every state. We verify the discounted best-response equations on every possible sample realization, including $\widehat{q} = 0$. Put $\lambda = 1 - \gamma$, let c be the maximum in (112), and set

$$X = \frac{c}{\lambda}, \qquad L = \gamma X, \qquad Y = X + \frac{1 - c}{\lambda + \gamma\widehat{q}}.$$

The hub rewards imply $0 < c \le 1/2$, so $L < X < Y$. Assign value X at x, value Y at y, and value L at every branch and padding state. Each empirical hub center puts mass $1 - t_{ab}$ on x and mass t_{ab} on branches of value L. Since L is the global minimum, full-simplex TV minimization gives

$$\phi_{\widehat{P}^0_{x,ab},r}(V) = (1 - \alpha_{ab})X + \alpha_{ab}L = X - c\alpha_{ab}.$$

Indeed, nature moves $\min\{r, 1 - t_{ab}\}$ mass from x to a branch. For every action pair, $R_{ab} \le c(1 + \gamma\alpha_{ab})$, with equality for the chosen pair. Rearranging this inequality gives

$$c = \max_{a,b}\{R_{ab} - \gamma c\alpha_{ab}\},$$

which is equivalent to the hub Bellman equation $X = \max_{a,b}\{R_{ab} + \gamma\phi_{\widehat{P}^0_{x,ab},r}(V)\}$. The chosen hub action therefore dominates every unilateral alternative. At a branch, either unilateral action against opponent action zero has zero reward and successor x, so its Bellman value is L. Padding states have the same conclusion. At y, the displayed formula for Y is equivalent to

$$Y = 1 + \gamma\{(1 - \widehat{q})Y + \widehat{q}X\}.$$

Since $Y > X$, the chosen joint action minimizes the exit probability and dominates every unilateral alternative at y. Thus V satisfies both the fixed-profile and each player's best-response Bellman equations. The standard discounted robust Bellman characterization in Proposition C.4 therefore proves exact empirical Nash optimality, including on samples with $\widehat{q} = 0$. A rule on other empirical kernels with the same known shared rewards and radii can use a deterministic centralized robust optimal policy whenever the stated support conditions fail. Such a policy is an exact Nash equilibrium because unilateral deviations are centralized strategies. Finite-action tie breaking gives a measurable extension.

3. Reduce the population gap to two scalar sample averages. With all branches low, a hub reward $R \in \{1/4, 1/4 + \eta\}$ gives robust average gain

$$\rho(R) = \frac{R}{1 + 2b + r}.$$

This is the calculation in Proposition I.2 with R substituted for $1/4 + c_N$. Its bias witness also verifies the value from every initial state. Against the low branch pattern, player 1's optimal response changes only its hub action to B; player 2 already maximizes its reward at the hub, and neither player can improve any branch or holding-state transition unilaterally in the population model. Consequently the population Nash gap is zero if the rule chooses B, and is exactly $\eta/(1 + 2b + r)$ if it chooses A.

If A is chosen, (112) gives

$$\eta(1 + \gamma\alpha_{A0}) \le \frac{\gamma}{4}(\alpha_{B0} - \alpha_{A0}).$$

The right side must be positive. Since $1+\gamma\alpha_{A0} \geq 1$, $\gamma < 1$, and $t \mapsto \min\{t+r, 1\}$ is nondecreasing and one-Lipschitz, we obtain $\eta \leq (t_{B0} - t_{A0})_+/4$. Dividing by $1 + 2b + r \geq 1$ gives

$$\mathrm{Gap}_{\mathrm{av}}(\mathfrak{O}_\gamma^{\mathrm{low}}(\widehat{P}^0)) \leq \frac{1}{4}(t_{B0} - t_{A0})_+. \tag{113}$$

If B is chosen, the same inequality holds because the Nash gap is zero. Both t_{A0} and t_{B0} are independent empirical averages of N Bernoulli$(2b)$ variables. Applying Hoeffding's inequality to their difference, viewed as a centered sum of $2N$ independent variables of range length $1/N$, yields

$$\Pr(t_{B0} - t_{A0} > 4\varepsilon) \leq \exp(-16N\varepsilon^2).$$

Together with (113), this proves (111). Since there are $4S$ primitive rows, taking $N = \max\{1, \lceil (16\varepsilon^2)^{-1} \log(1/\delta) \rceil\}$ proves the total sample bound. Positive reward rescaling by a factor at most one preserves the oracle and exactness and only reduces its population Nash gap. □

I.2.3 Centralized selection for shared-ambiguity team games

Proposition I.4. *Suppose all players have the same reward and the same primitive ambiguity sets, and a known H satisfies $H_{\mathrm{R}} \leq H < \infty$. The robust control problem with joint actions as its action set has a constant-gain solution with minimum bias span at most $2H$. Every deterministic stationary policy that is ε-optimal for this centralized problem from every state is a stationary product ε-Nash equilibrium of the game. If, in addition, every stationary centralized policy and every admissible stationary kernel induce a unichain, the robust-control learner of (Yang et al., 2026, Theorem 2 and the robust-only variant immediately following it) gives a designed learner with total sample complexity*

$$\widetilde{O}\left(SA_\times \frac{H + \sigma H^2}{\varepsilon^2}\right)$$

for $0 < \varepsilon \leq 1$, at fixed confidence, for full-simplex rectangular TV ambiguity with radii at most $\sigma \leq 1/2$.

Proof. Let w_γ be the centralized discounted optimal value. Its Bellman operator maximizes over the finite set of joint actions. Choose a deterministic maximizing joint policy $\boldsymbol{a}^\gamma$ at w_γ, and fix any player i. Against deterministic opponents $\theta^\gamma = \boldsymbol{a}^\gamma_{-i}$, the responding player's action maximum is bounded above by the centralized maximum. The particular action a_i^γ attains that maximum. Consequently

$$T_{i,\gamma}^{\theta^\gamma} w_\gamma = w_\gamma.$$

Contraction uniqueness identifies w_γ with this fixed-opponent optimal-response value.

By Theorem C.9, every such response value has span at most $2H$. Thus $\mathrm{sp}(w_\gamma) \leq 2H$ uniformly over discounts, even though the chosen opponents depend on γ. Lemma F.1, applied to the centralized finite-action robust MDP, supplies a constant-gain solution with span at most $2H$.

Write ρ_C^* for the centralized optimal robust gain. Every unilateral strategy against a fixed stationary opponent is an admissible centralized joint strategy with exactly the same reward and nature choices. Its robust payoff is therefore at most ρ_C^*. A deterministic centralized policy is a product policy because its selected joint action has deterministic components. If its gain is at least $\rho_C^* - \varepsilon$ from every state, each unilateral gain improvement is at most ε, proving the claimed Nash property.

For the statistical conclusion we invoke the cited robust-only control upper theorem, retaining its all-policy, all-kernel unichain assumption. The centralized MDP has S states, $A_\times$ actions, the same nominal samples and radii, and robust optimal bias-span parameter at most $2H$. Use the robust branch with the upper span bound $B = 2H$ and $\gamma = 1 - \varepsilon/(3B)$, returning a deterministic empirical discounted optimum. The proof of the cited theorem uses its span parameter only as an upper bound: its discounted Corollary 3 applies with reference span $2B$ and discounted accuracy B. Substitution gives $\widetilde{O}(SA_\times(B + \sigma B^2)/\varepsilon^2)$ samples. The unichain hypothesis also ensures fixed-policy robust Bellman solvability. The verification in Proposition C.6, applied with a singleton controller action, identifies the cited stationary-kernel gain with our history-dependent-nature criterion. Thus its policy guarantee transfers to the game by the preceding comparison. Exact empirical centralized optimality additionally makes this output an exact empirical discounted Nash oracle. □

The all-policy unichain hypothesis in the statistical part is satisfied by the joint lower-bound construction: every state outside x reaches x almost surely under every policy and kernel, so every closed recurrent class must contain x, and hence there is exactly one. Thus a successor-dimensional unrestricted minimax lower bound cannot hold throughout this shared-ambiguity unichain team subclass. It remains possible for the larger general-sum or player-specific-ambiguity class.

I.3 Continuation dependence and the support factor

The simultaneous concentration event in Section G pays for the support of each primitive nominal row. This subsection explains the continuation-query geometry and the dependence between a stationary empirical equilibrium and its transition samples.

I.3.1 The exact cost of a fully adaptive continuation query

The following identity applies to all continuation vectors, including vectors selected using the same row samples.

Proposition I.5. *Let $p \in \Delta([S])$, let $\widehat{p}$ be the empirical distribution from N independent samples, and define $\Delta = \widehat{p} - p$. For $\nu > 0$, let*

$$\mathcal{V}_\nu = \{v \in \mathbb{R}^S : \mathrm{Var}_p(v) \le \nu\}.$$

Then, almost surely,

$$\sup_{v \in \mathcal{V}_\nu} |\Delta^\top v| = \sqrt{\nu \sum_{x:p(x)>0} \frac{\Delta(x)^2}{p(x)}}. \tag{114}$$

Moreover,

$$\mathbb{E}\left[\sum_{x:p(x)>0} \frac{\Delta(x)^2}{p(x)}\right] = \frac{d(p)-1}{N}. \tag{115}$$

Proof. Almost surely $\widehat{p} \ll p$. Apply the variational representation of Pearson divergence, (Birrell et al., 2022, equation (15)), with $P = p$ and $Q = \widehat{p}$:

$$\chi^2(\widehat{p}\|p) = \sup_{\mathrm{Var}_p(f)>0} \frac{|(\widehat{p}-p)^\top f|^2}{\mathrm{Var}_p(f)}.$$

On the finite support all functions are integrable. Adding constants removes the reference's nonnegativity restriction, and scaling to variance ν proves (114). If $\widehat{p} = p$, including when $d(p) = 1$, the identity is immediate. Finally, the binomial marginal variance gives

$$\mathbb{E}\chi^2(\widehat{p}\|p) = \sum_{x:p(x)>0} \frac{p(x)(1-p(x))/N}{p(x)} = \frac{d(p)-1}{N},$$

which proves (115). □

Let $Z_\nu = \sup_{v \in \mathcal{V}_\nu} |\Delta^\top v|$. Squaring (114) and using (115) gives the exact identity $\mathbb{E}Z_\nu^2 = \nu(d(p) - 1)/N$. Thus the root-mean-square size of the fully adaptive query is $\sqrt{\nu(d(p)-1)/N}$. If $d(p) = 1$, then $\widehat{p} = p$ almost surely and this quantity is zero. The second-moment identity alone does not give a high-probability lower bound uniformly over all nominal rows. Nor does it give an equilibrium-learning lower bound, because an arbitrary vector in $\mathcal{V}_\nu$ need not be an equilibrium continuation value. The constructions in Theorems H.6 and H.12 address that additional requirement, with the latter also realizing the joint span–ambiguity dependence.

I.3.2 Temporal independence and stationary dependence

With independently sampled stage-specific kernels in a finite-horizon model, backward induction provides a direct conditional independence argument. Let $\mathcal{D}_h$ contain the samples for stage h, and let $\mathcal{F}_{h+1}$ be generated by all later-stage samples and any independent solver randomness. If the

continuation vector $\widehat{V}_{h+1}$ is $\mathcal{F}_{h+1}$-measurable, then, conditional on $\mathcal{F}_{h+1}$, each row in $\mathcal{D}_h$ is sampled independently of that vector. Scalar Bernstein concentration applies to a fixed clipping threshold. A deterministic threshold discretization, followed by a union bound and continuity of clipping, extends the bound to all thresholds. This explains the temporal decoupling used in finite-horizon robust-game analyses such as (Shi et al., 2024b). The stage-h equilibrium may depend on $\mathcal{D}_h$, since the later continuation vector has already been fixed.

A stationary empirical equilibrium does not have this stagewise construction. Its policy and continuation values solve coupled Bellman and best-response equations using the same transition rows. Consequently, freezing the empirical opponents in notation does not make them statistically independent of those rows. A theorem proved for one fixed response MDP needs a simultaneous event, a valid conditional independence argument, or another explicit stability argument before it can be applied at sample-selected opponents.

I.3.3 What single-agent and zero-sum results supply

The discounted optimal Bellman operator of a finite MDP is a contraction and has a unique optimal value vector. Absorbing-state constructions exploit this property to remove selected row samples from an auxiliary value calculation (Agarwal et al., 2020). In an ordinary two-player zero-sum discounted stochastic game, the Shapley operator likewise has a unique fixed-point value, even when optimal policies are nonunique (Zhang et al., 2023). These value-level uniqueness properties help address sample dependence in their respective models.

A general-sum equilibrium correspondence can contain several profiles with different continuation values. An arbitrary exact oracle may choose among these profiles using the row that is being estimated. The single-agent argument therefore does not transfer merely because a response problem becomes an MDP once opponents are fixed. In a robust model, one must also check that the zero-sum formulation preserves one common saddle-value operator. Player-specific worst-case evaluations do not provide that property in general.

I.3.4 The separate role of the oracle lower bound

Proposition I.5 describes unrestricted continuation queries. Theorems H.6 and H.12 supply the additional equilibrium realizability: their sample-dependent branch choices are exact empirical Nash equilibria, and the induced fluctuations change population incentives. Proposition I.3 shows why this implication depends on allowing arbitrary selection. Fixing the low branch collapses the relevant continuation query to a scalar branch-mass estimate and removes the support dependence on that family.

I.4 Additional requirements for alternative statistical methods

Method	Property it supplies	Remaining requirement in stationary general-sum games
Stagewise sampling	Independence from later-stage values	A stationary construction with a comparable independence property
Absorbing-state arguments	An auxiliary value that omits selected samples	Matching the chosen equilibrium branch after the perturbation
Equilibrium regularity	Local control of a specified solution branch	Uniform quantitative stability and concentration bounds
Entropy or logit responses	A unique regularized response to fixed opponents	Equilibrium uniqueness or branch control, and approximation error
Set-valued methods	Retention of multiple equilibrium payoffs	Concentration for the resulting value set
Independent validation	A candidate independent of validation samples	Two-sided value bounds and a candidate-generation guarantee
Uniform adaptive Bernstein	One event for all queried vectors	The deterministic propagation analysis of Section G

Table 1: What the methods provide and what a stationary general-sum analysis must still prove.

Regular equilibrium branches. Even if a specified equilibrium system admits a locally single-valued Lipschitz branch, this supplies only a local analytic property. A state-linear learning theorem additionally needs a neighborhood containing the random model perturbation, a controlled Lipschitz modulus, a compatible branch selection, and a concentration argument with the required dependence on dimension and discount. Those statistical conclusions do not follow from the local Lipschitz property alone.

Smoothing. Uniqueness of each player's entropy-regularized best response does not imply uniqueness of a regularized equilibrium. For example, in the two-player coordination game with payoff one for matching actions and zero for a mismatch, a symmetric logit equilibrium at temperature τ satisfies

$$p = f_\tau(p), \qquad f_\tau(p) = \frac{1}{1 + \exp(-(2p-1)/\tau)}.$$

The midpoint $p = 1/2$ is a solution. Define $g_\tau(p) = f_\tau(p) - p$. Then $g_\tau(0) > 0$, $g_\tau(1) < 0$, and $g'_\tau(1/2) = 1/(2\tau) - 1 > 0$ for $0 < \tau < 1/2$. Differentiability makes g_τ negative immediately to the left of $1/2$ and positive immediately to its right. The intermediate value theorem therefore gives one further root in $(0, 1/2)$ and one in $(1/2, 1)$. Thus at least three regularized equilibria exist even though each regularized best response is unique. Any smoothing-based statistical theorem must separately control equilibrium selection and the error relative to the original Nash criterion.

Set-valued methods. Retaining all equilibrium payoffs avoids an arbitrary choice during an existence argument. For learning, however, one must control the actual set of candidate continuation vectors. Proposition I.5 applies if that set is enlarged to the full variance ball. It does not establish that the equilibrium value set itself has the same dimension, nor exclude a more economical certificate based on its structure.

Independent validation. If a profile is constructed from one batch, then conditional on that batch the profile is fixed relative to an independent validation batch. This removes the dependence of the candidate on validation samples. It does not by itself bound the candidate's robust value or its best-response value. A near-optimal-policy theorem is also not automatically a two-sided value-estimation theorem. To certify a Nash gap, one needs valid upper bounds on best-response values and lower bounds on candidate values, with the necessary span or other regularity assumptions explicitly verified. An end-to-end learning rate further requires a guarantee that the candidate-generation procedure produces a profile that passes such validation.

I.4.1 Consequences for future improvements

Improving an upper bound without changing its oracle quantifier must respect the corresponding lower construction. A different oracle may instead restrict which continuation vectors can depend on a given row. Potential approaches include proving a quantitative branch-stability theorem, generating candidates followed by independent value certification, or constructing direct Bellman inequalities that certify deviations. Each is a separate research objective until the required statistical and algorithmic steps are established. No state-linear theorem follows solely from naming one of these mechanisms.

Bounded nominal support and factored transition structure should also be distinguished. A locally factored model may still have a nominal row supported on the entire global state space. Such a model only falls within a bounded-support result when its actual successor support is bounded, or when a separate argument exploits the factorization.