VECTOR BELLMAN THEORY FOR MULTICHAIN ROBUST AVERAGE-REWARD MARKOV DECISION PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust average-reward Markov decision processes provide a fundamental framework for long-term performance optimization under uncertainty, and can have optimal long-run rewards that depend on the initial state. This state dependence requires a vector Bellman theory that accounts for both recurrent-class rewards and transition uncertainty. We develop such a theory for finite models with compact, post-action (s, a)-rectangular ambiguity. A gain-first, bias-second optimization principle yields a coupled vector gain-bias system, and every finite solution identifies the optimal robust gain and supplies stationary saddle strategies against history-dependent opponents, simultaneously from all initial states. We further characterize solvability through stationary gain conditions and a uniform bound on canonical transient corrections, and give sufficient conditions that permit distinct recurrent-class gains. The certificates also yield asymptotically affine trajectories of the robust Bellman operator, based on which we design a robust approximately shifted Halpern planning algorithm. Under finite Bellman solvability, the gain estimates and Bellman displacements converge to the optimal gain vector, and every extracted greedy controller is average-optimal after a finite, instance-dependent budget. These results thus connect finite Bellman certificates to undiscounted planning for state-dependent robust average rewards, providing theoretical understandings.

1 INTRODUCTION

Markov decision processes (MDPs) (Puterman, 2014) provide a standard framework for sequential decision-making under stochastic agent-environment interactions, in which an agent seeks a policy that maximizes expected reward under a specified performance criterion. A policy optimized for a single nominal MDP with a fixed transition model, however, can perform poorly when the deployed dynamics differ from that model or are imperfectly known. Such model mismatches widely exist in practice, due to, e.g., inaccurate model estimation, non-stationary environment, or unexpected perturbations upon deployment. Robust MDPs are thus proposed to address this mismatch by optimizing worst-case performance over a prescribed family of transition models (Iyengar, 2005; Nilim & El Ghaoui, 2004), which captures the uncertain models, and is named ambiguity set or uncertainty set. To find such a robust policy, one need to set the performance criterion. Among different criteria, average reward is particularly important as it measures sustained performance without imposing a discount factor, making it natural for systems operating over long horizons. Robust average-reward theory must therefore reconcile uncertainty in the transition dynamics with the chain structure that governs long-run performance.

In average-reward models, the limiting average-reward function (i.e., the long-run gain) is naturally dependent on the initial state. Even under a fixed stationary policy and fixed transition kernel, different starting states may reach recurrent classes with different class gains and different absorption probabilities: even two absorbing states with unequal rewards can yield a nonconstant gain vector. Under transition uncertainty, stationary gains can also change abruptly and becomes more challenging. If a zero-reward state reaches an absorbing unit-reward state with probability ε per step, its gain is 1 for every $\varepsilon > 0$ but 0 at $\varepsilon = 0$. Vanishing transition probabilities can therefore change the recurrent structure. This raises two separate questions: whether nature attains the worst stationary gain and whether transient rewards admit a finite common bias. These features thus require a state-dependent gain vector and an all-state gain-bias formulation, to tackle the complicated statewise dependence and provide a concrete Bellman-typed characterization.

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.

Existing work supplies two complementary foundations for such a theory. The early line of work (Wang et al., 2023d; 2024d) develop the robust Bellman equation and algorithms for (s, a)-rectangular models under a uniform unichain condition, which ensures all the average reward (or the gain) is independent from initial state. More recently, Wang & Si (2025) develop robust Bellman optimality with constant (optimal) gain, which covers more general conditions like one-sided weak communication. These conditions may permit individual stationary controller-nature pairs to be multichain, while the Bellman result remains in the state-independent-gain regime. Strategic value theory developed in (Grand-Clement et al., 2023) establishes stationary controller optimality for general compact (s, a)-rectangular ambiguity, while polytopic models admit planning through finite stochastic games (Chatterjee et al., 2024). But no Bellman-typed characterization is studied. These studies thus leave an open question: *when the gain depends on the initial states, does this vector gain admit a finite gain-bias Bellman certificate, and how can such a certificate support planning?* In this work, we provide concrete answers to the question. Our main contributions are summarized as follows.

A gain-first, bias-second vector Bellman certificate. We first formulate a coupled Bellman system that gives continuation gain priority for both players (nature and controller) to characterize the robust average reward. Nature first minimizes continuation gain, and the controller maximizes the resulting minimum. Reward-bias optimization then takes place among the choices tied at this first level. We further show that every finite solution identifies the robust optimal gain and produces an all-state stationary saddle against history-dependent opponents. This verification controls nature's deviations from gain-minimizing rows, including rows with arbitrarily small gain gaps. It also connects the certificate to an asymptotically affine trajectory of the original Bellman operator (Section 4).

Exact solvability beyond communication assumptions. We then characterize when the finite certificate exists. For optimal control, the criterion combines a stationary saddle in the gain-restricted game with a uniform lower bound on canonical biases over all zero-gain replies to one securing controller. This separates stationary gain attainment from the control of transient rewards needed for a finite bias. We further explore concrete regimes that permit distinct recurrent-class gains: polytopic row sets, a uniform support gap, and different concrete distributional ambiguity sets, showing they are sufficient conditions for Bellman solvability (Section 5).

Direct undiscounted planning with a vanishing affine defect. We adapt approximately shifted Halpern iteration (Zurek & Chen, 2025, Algorithm 1) to the robust Bellman operator. For compact row sets, a finite Bellman certificate can produce an asymptotically affine trajectory whose error remains nonzero at every finite time. We show that this vanishing defect suffices for convergence of the gain estimates and Bellman displacements. Directional control of the iterates then makes every extracted greedy controller eventually average-optimal from all states. The analysis gives finite-budget bounds in terms of the affine defect and compact examples with arbitrarily small algebraic convergence exponents for the gain estimator (Section 6).

Roadmap. Section 3 establishes the robust gain vector and an all-state optimal stationary controller under the standing assumptions. Section 4 develop the vector Bellman equation system and optimality guarantees of its solutions. Section 5 characterizes when that solution exists and identifies sufficient regimes. Section 6 uses the certificate to analyze direct undiscounted planning. Thus the gain and optimal controller precede the Bellman system logically; finite solvability is the condition for the certificate and the planning theorem.

2 Problem Formulation and Preliminaries

We consider a finite robust MDP $(S, A, \mathcal{U}, r)$. The state space is $S = \{1, \ldots, n\}$, each action set $A(i)$ is finite and nonempty, and $r_{ia} \in \mathbb{R}$ is the one-stage reward. At time $t \geq 0$, the controller observes S_t and chooses $A_t \in A(S_t)$. Nature observes the chosen action, selects a transition row $p_t \in \mathcal{U}_{S_t A_t} \subseteq \Delta(S)$, and the next state is sampled from p_t. We assume (s, a)-rectangular ambiguity (Iyengar, 2005; Nilim & El Ghaoui, 2004): $\mathcal{U} = \prod_{i \in S} \prod_{a \in A(i)} \mathcal{U}_{ia}$, where every row set $\mathcal{U}_{ia}$ is nonempty and compact. In the following, vector inequalities are componentwise.

The planning objective is to compute a controller policy that maximizes worst-case long-run reward. Write $\Pi_D \subseteq \Pi_S \subseteq \Pi_H$ for deterministic stationary, randomized stationary, and randomized history-dependent controller strategies. A stationary policy belongs to $\prod_i \Delta(A(i))$, while a history-dependent policy maps the observed history $(S_0, A_0, \ldots, S_t)$ to a distribution on $A(S_t)$. Nature's randomized

history-dependent strategies form $\mathcal{Q}_H$, and its full stationary selectors form $\mathcal{Q}_S = \prod_i \prod_{a \in A(i)} \mathcal{U}_{ia}$, is the transition kernels nature selected. A full selector specifies a row for every state-action pair, including actions unused by a particular controller. Both players observe the initial state.

Payoffs. Our primary performance criterion is the lower limiting expected average reward. We also record the corresponding upper limit and the two criteria that take the sample-path limit before expectation. These distinctions specify the strength of the strategy guarantees proved below. For $(\sigma, \tau) \in \Pi_H \times \mathcal{Q}_H$, let $\mathbb{E}_i^{\sigma,\tau}$ denote expectation under the induced process starting at $S_0 = i$. For the trajectory $(S_0, A_0, S_1, A_1, \ldots)$, define $X_N = N^{-1} \sum_{t=0}^{N-1} r_{S_t A_t}$ and $J_N(i; \sigma, \tau) = \mathbb{E}_i^{\sigma,\tau} X_N$. We distinguish four average-payoff conventions (Puterman, 2014):

$$\begin{aligned} J_i^- &= \liminf_N \mathbb{E}_i X_N, \quad J_i^+ = \limsup_N \mathbb{E}_i X_N, \\ I_i^- &= \mathbb{E}_i \liminf_N X_N, \quad I_i^+ = \mathbb{E}_i \limsup_N X_N, \end{aligned} \tag{1}$$

with the strategy pair suppressed. These criteria can differ for a fixed history-dependent pair. For a payoff Ψ and strategy classes $\mathcal{C} \subseteq \Pi_H, \mathcal{N} \subseteq \mathcal{Q}_H$, define the lower and upper values

$$\underline{v}_i^{\Psi}(\mathcal{C}, \mathcal{N}) := \sup_{\sigma \in \mathcal{C}} \inf_{\tau \in \mathcal{N}} \Psi_i(\sigma, \tau), \qquad \overline{v}_i^{\Psi}(\mathcal{C}, \mathcal{N}) := \inf_{\tau \in \mathcal{N}} \sup_{\sigma \in \mathcal{C}} \Psi_i(\sigma, \tau). \tag{2}$$

For a fixed controller σ, its robust performance is $\inf_{\tau \in \mathcal{Q}_H} \Psi_i(\sigma, \tau)$. The lower value maximizes this guarantee, while the upper value minimizes the controller's best response. A pair $(\bar{\sigma}, \bar{\tau})$ is an all-state saddle for Ψ at $u \in \mathbb{R}^S$ if $\Psi_i(\bar{\sigma}, \tau) \geq u_i \geq \Psi_i(\sigma, \bar{\tau})$ for every i, $\sigma \in \Pi_H$, and $\tau \in \mathcal{Q}_H$. Theorem F.1 identifies a common value for the four conventions, and Theorem H.1 characterizes exact stationary attainment by nature.

Dynamic programming operators. The order of play within each stage explains the robust Bellman maps. Given continuation value x at state i, the controller first chooses a and nature, having observed that action, chooses $p \in \mathcal{U}_{ia}$. The resulting one-step saddle operators are (Iyengar, 2005)

$$(Tx)_i = \max_{a \in A(i)} \left\{ r_{ia} + \min_{p \in \mathcal{U}_{ia}} p^\top x \right\}, \qquad (T^\pi x)_i = r_i^\pi + \sum_a \pi(a \mid i) \min_{p \in \mathcal{U}_{ia}} p^\top x,$$

where $r_i^\pi = \sum_a \pi(a \mid i) r_{ia}$ for $\pi \in \Pi_S$. Nature observes the sampled action and then selects the row, which explains the separate minima in T^π. For discount factor $1 - \epsilon$, $0 < \epsilon < 1$, define the value vectors coordinatewise by $(V_\epsilon^\pi)_i = \inf_{\tau \in \mathcal{Q}_H} \mathbb{E}_i^{\pi,\tau} \sum_{t=0}^{\infty} (1-\epsilon)^t r_{S_t A_t}$, and $(V_\epsilon)_i = \sup_{\sigma \in \Pi_H} \inf_{\tau \in \mathcal{Q}_H} \mathbb{E}_i^{\sigma,\tau} \sum_{t=0}^{\infty} (1-\epsilon)^t r_{S_t A_t}$. Rectangular discounted dynamic programming (Iyengar, 2005; Nilim & El Ghaoui, 2004) gives $V_\epsilon = T((1-\epsilon)V_\epsilon)$ and $V_\epsilon^\pi = T^\pi((1-\epsilon)V_\epsilon^\pi)$. Each discounted fixed point is unique and is attained simultaneously from all states by deterministic stationary discounted-optimal selectors. Consequently, $(V_\epsilon)_i = \max_{\pi \in \Pi_D} (V_\epsilon^\pi)_i$. Appendix C reviews the discounted dynamic-programming and finite-chain facts used in the analysis.

Multichain structure and vector gain. A stationary pair (π, q) induces the transition matrix $P_{ij}^{\pi,q} = \sum_a \pi(a|i) q_{ia,j}$. Its gain vector is $\eta^{\pi,q} := (P^{\pi,q})^\infty r^\pi$, where the Cesàro projector $P^\infty := \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} P^t$ exists for every finite stochastic matrix, including periodic multichain matrices, and each coordinate of $\eta^{\pi,q}$ is an absorption-weighted average of recurrent-class rewards (Puterman, 2014). This representation explains how initial states can have different gains. Unichain and irreducible assumptions in prior analyses make each stationary chain's gain constant (Wang et al., 2023d; 2024d; Xu et al., 2025a; Roch et al., 2025). Here both evaluation and control concern the complete statewise gain vector. See Appendix C for a review of existing results.

3 Statewise values and stationary strategies

We first identify the gain vector, the main objective of our studies. The results connect discounted and finite-horizon values to robust average reward and supply strategies that secure these values simultaneously from every initial state. We begin with a fixed stationary controller.

Theorem 3.1. *[Extension of Grand-Clement et al. (2023, Lemmas 3.3 and 4.7)] For every $\pi \in \Pi_S$, there is $g^\pi \in \mathbb{R}^S$ such that, with both limits in supremum norm,*

$$g^\pi = \lim_{\epsilon \downarrow 0} \epsilon V_\epsilon^\pi = \lim_{N \to \infty} \frac{(T^\pi)^N 0}{N}, \text{ and } g_i^\pi = \inf_{q \in \mathcal{Q}_S} \eta_i^{\pi,q}, \forall i \in S. \tag{3}$$

For every $\Psi \in \{I^-, J^-, J^+, I^+\}$, $\inf_{\tau \in \mathcal{Q}_H} \Psi_i(\pi, \tau) = \inf_{q \in \mathcal{Q}_S} \Psi_i(\pi, q) = g_i^\pi$. *Moreover, for each* $\nu > 0$, *there exists some* $q^{\pi,\nu} \in \mathcal{Q}_S$ *such that* $g^\pi \leq \eta^{\pi, q^{\pi,\nu}} \leq g^\pi + \nu \mathbf{1}$.

The vector g^π is the robust gain of π: it is the common normalized discounted and finite-horizon limit and the worst-case value under each payoff convention in equation 1. A stationary selector attaining this vector need not exist, even though the row sets are compact. For every positive tolerance, however, one stationary selector approximates the entire vector simultaneously. This is the all-state guarantee used in policy evaluation.

Remark 1. *Grand-Clement et al. (2023, Lemmas 3.3 and 4.7) establish stationary evaluation and the normalized discounted limit in equation 3. We additionally obtain the finite-horizon limit and formulate simultaneous approximation by one selector.*

The next result identifies the optimal robust gain and strategies securing it from all initial states.

Theorem 3.2. *[Extension of Grand-Clement et al. (2023, Theorem 5.2)] There exist* $g^\star \in \mathbb{R}^S$ *and a deterministic stationary* $\pi^\star \in \Pi_D$ *such that*

$$g^\star = \lim_{\epsilon \downarrow 0} \epsilon V_\epsilon = \lim_{N \to \infty} \frac{T^N 0}{N} = \max_{\pi \in \Pi_D} g^\pi = g^{\pi^\star}, \tag{4}$$

where the limits are in supremum norm and the maximum is coordinatewise. Moreover, for every $\delta > 0$, *there exist* $q_\delta \in \mathcal{Q}_S$ *and* $N_\delta < \infty$ *such that, for all* $i \in S$ *and* $N \geq N_\delta$,

$$J_N(i; \pi^\star, \tau) \geq g_i^\star - \delta \text{ for every } \tau \in \mathcal{Q}_H, \text{ and } J_N(i; \sigma, q_\delta) \leq g_i^\star + \delta \text{ for every } \sigma \in \Pi_H. \tag{5}$$

Equation 4 identifies the optimal robust gain through discounted values, finite-horizon values, and policy optimization. One deterministic stationary controller $\pi^\star$ attains every coordinate. For each tolerance δ, equation 5 supplies a single stationary nature selector and one horizon threshold that work for every initial state, every later horizon, and every history-dependent opponent. The controller remains the same for all tolerances; nature's approximating selector can depend on δ.

Remark 2. *Convergence of* $T^N 0/N$ *holds for compact row sets without any additional assumption. Theorem 3.2 therefore supplies a planning target throughout the model class. Later we will study the stronger certificate and the guarantees that follow when that certificate is finite. Our finite-horizon conclusion extends Grand-Clement et al. (2023, Theorem 5.2) beyond definable ambiguity.*

The vector $g^\star$ is also the common lower and upper value under all four payoff conventions. The same controller $\pi^\star$ is optimal in each case. The values are preserved for intermediate strategy classes containing Π_D and $\mathcal{Q}_S$, respectively, including the stationary strategy classes. See Appendix F.

For a full stationary selector q, let $d_i(q) := \max_{\pi \in \Pi_D} \eta_i^{\pi,q}$ be the optimal gain of the nominal MDP with q fixed. Nature is exactly optimal from all states against every history-dependent controller if and only if $d(q) = g^\star$. This criterion checks the controller's best response over every action, including actions unused by $\pi^\star$. The equality $\eta^{\pi^\star, q} = g^\star$ checks only the prescribed controller. Appendix H develops this characterization directly from stationary gains.

4 Vector Bellman equation system

Theorems 3.1-3.2 identify the robust value and an all-state optimal controller without assuming a finite bias. In this section, we investigate a stronger certificate: a gain-bias certificate: it must reconcile long-run gains and transient rewards for both players using one finite pair of vectors. We now formulate a local system and establish what any finite solution certifies.

4.1 Robust Bellman optimality system: Gain first, bias second

In the constant-gain setting (Wang et al., 2023d; Wang & Si, 2025), the robust Bellman equation is $\rho \mathbf{1} + h = Th$. Every probability row preserves the same continuation gain because $p^\top(\rho \mathbf{1}) = \rho$. For a vector gain, different rows can lead to different long-run reward rates. The equation $g + h = Th$ alone does not require the selected transitions to preserve the proposed gain. The Bellman system must therefore compare continuation gains before comparing finite reward and bias terms.

To see the required order, consider a continuation vector $tg + h$ for large t. The one-step objective is $tp^\top g + r_{ia} + p^\top h$. A fixed gain difference dominates the bounded reward-bias term as t grows. Nature therefore first minimizes $p^\top g$, and the controller maximizes this minimum. Among the choices tied in gain, both players optimize reward and continuation bias. Our system below imposes this ordering on both players, extending the nominal multichain gain-bias separation (Puterman, 2014; Zurek & Chen, 2025).

Gain first, bias second. For $g \in \mathbb{R}^n$, define the worst continuation gain of each action and the controller's optimal continuation gain by

$$m_{ia}(g) := \min_{p \in \mathcal{U}_{ia}} p^\top g, \qquad (\widehat{T}g)_i := \max_{a \in A(i)} m_{ia}(g). \tag{6}$$

For any $g \in \mathbb{R}^n$, define

$$A_g(i) := \{a \in A(i) : m_{ia}(g) = g_i\}, \qquad F_{ia}(g) := \arg\min_{p \in \mathcal{U}_{ia}} p^\top g, \tag{7}$$

and we call $F_{ia}(g)$ a gain face even when the row set is nonconvex. Compactness makes every $F_{ia}(g)$ nonempty. When $g = \widehat{T}g$, each $A_g(i)$ is also nonempty, and we define the optimal-control bias operator on these active actions. Moreover, define the bias operator

$$(L_g h)_i := \max_{a \in A_g(i)} \min_{p \in F_{ia}(g)} (r_{ia} + p^\top h), \qquad K_g h := L_g h - g. \tag{8}$$

Our robust vector Bellman optimality system is

$$g = \widehat{T}g, \qquad g + h = L_g h. \tag{9}$$

The two equations perform distinct tasks. The first enforces consistency of continuation gain, while the second determines the reward-bias balance among gain-optimal choices. They must be solved together: the first equation is reward-free and accepts every constant vector. The restrictions to $A_g(i)$ and $F_{ia}(g)$ preserve gain priority for the controller and nature, respectively. Specifically, Examples 1-2 show that omitting either restriction can certify an incorrect gain.

For singleton row sets, the system reduces to the classical multichain MDP optimality equations (Puterman, 2014). If $g = \rho\mathbf{1}$, all actions and rows are gain-active, so $L_g = T$ and equation 9 becomes the constant-gain robust Bellman equation (Wang et al., 2023d; Wang & Si, 2025).

The following theorem verifies the system against history-dependent opponents.

Theorem 4.1 (Bellman optimality)**.** *Every finite solution (g, h) of equation 9 satisfies $g = g^\star$. Choose a deterministic stationary policy $\pi(i) \in \operatorname{argmax}_{a \in A_g(i)} \min_{p \in F_{ia}(g)} \{r_{ia} + p^\top h\}$, and for every active state-action pair choose $q_{ia} \in \arg\min_{p \in F_{ia}(g)} p^\top h$; at inactive actions choose any $q_{ia} \in F_{ia}(g)$. Then (π, q) forms an all-state saddle for every payoff in equation 1:*

$$\liminf_N J_N(i; \pi, \tau) \geq g_i \quad \forall \tau \in \mathcal{Q}_H, \qquad \limsup_N J_N(i; \sigma, q) \leq g_i \quad \forall \sigma \in \Pi_H, \tag{10}$$

and this pair achieves the optimal robust average reward: $\lim_{N\to\infty} J_N(i; \pi, q) = g_i^\star$ for every $i \in S$.

Fix the selected controller and write $d(p) = p^\top g - g_i \geq 0$ and $e(p) = r_{ia} + p^\top h - g_i - h_i$. On a gain-minimizing row, the bias equation gives $e(p) \geq 0$. Compactness implies that for each $\theta > 0$ there is a finite C_θ with $e(p) \geq -\theta - C_\theta d(p)$ on every feasible row. Along any history-dependent nature strategy, the cumulative expected gain increase is bounded by $\operatorname{sp}(g)$; telescoping the bias gives the controller's lower guarantee. For the selected full nature plan, active controller actions satisfy the reverse bias inequality, while inactive actions have fixed negative continuation-gain gaps. Since there are finitely many controller actions, their bias discrepancies can be charged to those gaps. The same telescoping argument gives nature's upper guarantee.

Asymptotically affine Bellman trajectory. The certificate also describes the behavior of the original Bellman operator T, which optimizes over all actions and all feasible rows. Our Lemma I.1 proves

$$g = \widehat{T}g, \qquad g + h = L_g h \quad \Longleftrightarrow \quad T(h + tg) = h + (t+1)g + o(1), \qquad t \to \infty.$$

This relation connects the gain-restricted system to the original operator used by the planner. As t grows, actions with a fixed continuation-gain disadvantage cease to compete, and minimizing rows approach their gain faces. An optimal row for $h + tg$ can nevertheless lie outside its gain face at every finite t. Section 6 therefore tracks a vanishing affine defect to design the planning algorithm.

Remark 3. *A finite bias imposes additional one-step compatibility beyond average optimality: a prescribed optimal controller or stationary saddle need not be certifiable by a common bias, even when another Bellman solution exists (Example 3). Biases also need not be unique after fixing one reference state, since the gain faces can preserve further harmonic directions.*

4.2 FIXED-POLICY EQUATION

For a fixed stationary controller, the Bellman system certifies its robust gain and a stationary nature selector that is worst from all initial states. The action maximum is replaced by the policy average, while nature continues to minimize separately after each realized action.

Proposition 4.2 (Fixed-policy Bellman equations)**.** *Fix any stationary randomized policy $\pi \in \Pi_S$. Suppose finite vectors (g, h) satisfy the fixed-policy Bellman system: for every state i,*

$$g_i = \sum_a \pi(a \mid i) \min_{p \in \mathcal{U}_{ia}} p^\top g, \quad g_i + h_i = r_i^\pi + \sum_a \pi(a \mid i) \min_{p \in F_{ia}(g)} p^\top h, \tag{11}$$

where $r_i^\pi := \sum_a \pi(a \mid i) r_{ia}$, and $F_{ia}(g) := \arg\min_{p \in \mathcal{U}_{ia}} p^\top g$. Then $g = g^\pi$. Moreover, for every state-action pair with $\pi(a \mid i) > 0$, choose $q_{ia}^\star \in \arg\min_{p \in F_{ia}(g)} p^\top h$, and choose arbitrary feasible rows for zero-probability actions. Then for every initial state i,

$$\inf_{\tau \in \mathcal{Q}_H} \liminf_{N \to \infty} J_N(i; \pi, \tau) = \inf_{\tau \in \mathcal{Q}_H} \limsup_{N \to \infty} J_N(i; \pi, \tau) = g_i^\pi, \tag{12}$$

and the single stationary full selector $q^\star$ attains both infima simultaneously at every initial state.

The same finite pair certifies the complete robust gain vector and one stationary selector attaining it. We will later characterize exactly when such a pair exists.

5 EXACT STATIONARY CONDITIONS FOR SOLVABILITY

Section 4 establishes that a finite Bellman solution guarantees the optimal robust gain and optimal controller. We now determine when such a solution exists.

A finite bias requires both attainment of the stationary gain and uniform control of the associated transient reward corrections. We first make these requirements precise for policy evaluation. We then characterize optimal-control solvability through a stationary saddle in the gain-restricted game and a one-sided bound on its canonical biases. For a finite stochastic matrix P, let $P^\infty = \lim_{N \to \infty} N^{-1} \sum_{t=0}^{N-1} P^t$ and $Z_P = (I - P + P^\infty)^{-1}$. Both exist without irreducibility or aperiodicity (Puterman, 2014). Fix $\pi \in \Pi_S$ and define its effective row set by $\mathcal{U}_i^\pi = \{\sum_{a \in A(i)} \pi(a \mid i) p_a : p_a \in \mathcal{U}_{ia} \text{ for every } a \in A(i)\}$. Independent actionwise minimization gives $(T^\pi v)_i = r_i^\pi + \min_{p \in \mathcal{U}_i^\pi} p^\top v$. Thus, nature's fixed-policy problem is a compact-action MDP with reward r^π. For an effective selector $q \in \prod_i \mathcal{U}_i^\pi$, write $(P_q)_{i\cdot} = q_i^\top$.

Theorem 5.1 (Fixed-policy solvability)**.** *Fix $\pi \in \Pi_S$. For effective selectors $q \in \prod_i \mathcal{U}_i^\pi$, define $\eta(q) := P_q^\infty r^\pi$, $\gamma_i^\pi := \inf_q \eta_i(q)$, $\mathcal{Q}_*^\pi := \{q : \eta(q) = \gamma^\pi\}$, and $w(q) := Z_{P_q}(r^\pi - \eta(q))$. Then equation 11 has a finite solution if and only if*

$$\mathcal{Q}_*^\pi \neq \varnothing \quad \text{and} \quad \exists B < \infty \text{ such that } w(q) \geq -B\mathbf{1} \text{ for any } q \in \mathcal{Q}_*^\pi. \tag{13}$$

Every finite solution (g, h) satisfies $g = g^\pi = \gamma^\pi$.

The two conditions identify separate requirements. The set Q_*^π contains the stationary kernels that attain the robust gain simultaneously from every state. For such a kernel, $w(q)$ is its canonical transient reward correction, with normalization $P_q^\infty w(q) = 0$. The uniform lower bound prevents these normalized corrections from becoming arbitrarily negative across gain-attaining kernels. Average-gain attainment and finite transient corrections are therefore distinct parts of Bellman solvability.

For optimal control, the candidate gain supplies the appropriate reward centering. Every gain-active controller-nature pair satisfies $P^{\pi,q} g = g$, so replacing r_{ia} by $c_{ia} = r_{ia} - g_i$ subtracts g from its original gain. The bias equation is therefore a zero-gain problem on the gain-restricted action and row sets. A stationary saddle secures this zero gain, and a bound on transient corrections determines

whether the saddle value has a finite Bellman representation. The active controller policies and nature selectors are $\Pi_g := \prod_i A_g(i)$, $\mathcal{Q}_g := \prod_i \prod_{a \in A_g(i)} F_{ia}(g)$, and $\mathcal{Q}_g^\pi := \prod_i F_{i,\pi(i)}(g)$. A full plan $q \in \mathcal{Q}_g$ specifies rows for every active action, while a reply $q \in \mathcal{Q}_g^\pi$ specifies only the rows used by π. Accordingly, $P_{\pi,q}$ has row $q_{i,\pi(i)}^\top$ for a full plan and row $q_i^\top$ for a restricted reply. Set $c_i^\pi = c_{i,\pi(i)}$, $\eta^{\pi,q} = P_{\pi,q}^\infty c^\pi$, and $w^{\pi,q} = Z_{P_{\pi,q}} c^\pi$ when $\eta^{\pi,q} = 0$. Here $\eta^{\pi,q}$ uses the centered rewards c, whereas $\eta(q)$ in Theorem 5.1 uses the original rewards r^π.

Theorem 5.2 (Optimal-control solvability)**.** *Fix $g \in \mathbb{R}^S$ with $g = \widehat{T} g$. There is a finite h with $K_g h = h$ if and only if there exist $\bar{\pi} \in \Pi_g$, a full plan $\bar{q} \in \mathcal{Q}_g$, and $B < \infty$ such that*

$$(i)\eta^{\bar{\pi},q} \geq 0, \text{ for every } q \in \mathcal{Q}_g^{\bar{\pi}}; \quad (ii)\eta^{\pi,\bar{q}} \leq 0, \text{ for every } \pi \in \Pi_g;$$
$$(iii) w^{\bar{\pi},q} \geq -B\mathbf{1}, \text{ for every } q \in \mathcal{Q}_g^{\bar{\pi}} \text{ with } \eta^{\bar{\pi},q} = 0.$$

Equivalently, finite solvability is equivalent to $\sup_{N \geq 0} \|K_g^N 0\|_\infty < \infty$. Another equivalent condition is the existence of finite ℓ, u with $\ell \leq K_g \ell$ and $K_g u \leq u$.

Conditions (i) and (ii) give a zero-gain saddle within the gain-restricted game: one controller secures nonnegative gain against every restricted nature reply, and one full nature plan holds every active controller to nonpositive gain. Condition (iii) supplies the remaining transient control. Each fixed zero-gain reply has a finite canonical bias, but the biases can lack a common lower bound as transition probabilities vanish and recurrent classes change. The bound is one-sided because, after fixing nature's full plan, finitely many deterministic controller policies remain. The theorem thus locates the gap between stationary gain optimality and a common finite Bellman bias. Appendix J further treats the case in which the gain is not prescribed.

A prescribed average-optimal controller or stationary saddle can fail to share a Bellman bias even when another pair supports a finite certificate (Example 3). Appendix K characterizes which active controller-nature pairs share one bias satisfying both players' Bellman inequalities. The characterization determines the exact minimum span for each compatible pair and, after optimization over pairs, for the full Bellman system. It also describes the remaining bias freedom through recurrent-class offsets, which can persist after fixing one reference state.

Concrete sufficient regimes for solvability. The following conditions ensure finite Bellman solvability while allowing different recurrent classes to have different gains (beyond constant gains). They provide concrete model classes covered by the preceding criterion and the planning result later.

Proposition 5.3 (Sufficient regimes for Bellman solvability)**.** *Under the standing assumptions, each of the following conditions ensures a finite solution $(g^\star, h)$ to equation 9:*

(A) `Polytopic ambiguity`: Every row set $\mathcal{U}_{ia}$ is a polytope.

(B) `Uniform support gap`: There exists $\delta > 0$ such that every feasible row satisfies $p_j = 0$ or $p_j \geq \delta$, for all i, a, $p \in \mathcal{U}_{ia}$, and $j \in S$.

(C) `Continuous stationary projections`: For every $\pi \in \Pi_D$, the map $P \mapsto P^\infty$ is continuous on the compact induced-kernel family $\mathcal{P}^\pi = \{P^{\pi,q} : q \in \mathcal{Q}_S\}$.

Polytopic rows make T piecewise affine, so Kohlberg's invariant half-line theorem gives a finite pair (Kohlberg, 1980). Under (C), stationary gains vary continuously with the kernel: compactness upgrades simultaneous stationary approximations to an exact full nature plan, while continuity of the Cesàro projections uniformly bounds the canonical corrections of zero-gain replies. These are the requirements in Theorem 5.2; condition (B) implies (C), as proved in Appendix M. These regimes accommodate a nonconstant optimal gain. In that case, $\epsilon \operatorname{sp}(V_\epsilon) \to \operatorname{sp}(g^\star) > 0$, so the uncentered discounted span grows as the discount vanishes. The existence analysis in Appendix M.1 controls the finite remainder $V_\epsilon - g^\star/\epsilon$ in these regimes. This vector centering isolates transient rewards while preserving the distinct long-run gains of recurrent classes. It extends the role played by bounded discounted span in constant-gain robust theory (Wang & Si, 2025, Theorem 4).

Remark 4 (Solvability under concrete distributional uncertainty sets)**.** *Our results also imply solvability under several distributional uncertainty sets that are extensively studied in robust RL with additional unichain assumption (Roch et al., 2025; 2026), defined as $\mathcal{U}_{ia} := \{q \in \Delta_S : D(q||p_{ia}^0) \leq \rho_{ia}\}$, where D is some probability divergence, like total variation or KL-divergence, p^0 is some nominal kernel, and ρ_{ia} is the radius. Since total-variation balls are polytopes, so Proposition 5.3(A)*

applies at every radius. For forward KL balls $D_{\mathrm{KL}}(p\|p^0_{ia}) \le \rho_{ia}$, condition (B) applies when $\rho_{ia} < \min_{j:p^0_{ia,j}>0} -\log(1-p^0_{ia,j})$ for every row (we further provide an unsolvable example when this condition fails). Support-restricted reverse KL balls satisfy condition (B) at every finite radius. Appendix M.1.1 proves these and extends to other divergences.

Further structure of Bellman certificates. Appendix K refines the existence result by characterizing which active controller-nature pairs share one bias satisfying both players' Bellman inequalities. An average-optimal pair can fail this compatibility test even when a different pair supports a finite certificate. The characterization determines the exact minimum span for each compatible pair and, after optimization over pairs, for the full Bellman system. It also describes the remaining bias freedom through recurrent-class offsets, which can persist after fixing one reference state.

Remark 5. *Although the robust gain and an all-state optimal stationary controller exist under our standing assumptions (as Section 3 proves), the vector Bellman system need not admit a finite solution. This is different from non-robust cases: general compact ambiguity can obstruct stationary worst-gain attainment or a uniform lower bound on canonical biases, as illustrated in Appendix M.2; In contrast, nominal MDPs with finite state and action spaces always admit solutions to their multichain gain-bias system (Puterman, 2014; Schweitzer, 1985). This nominal solvability also follows from Proposition 5.3(A), since every nominal transition row corresponds to a singleton ambiguity set, thus a polytope. Our vector Bellman system thus provides a finite certificate with an exact solvability criterion of the multichain robust average-reward MDPs, generalizing the non-robust Bellman theory.*

6 Planning from Finite Bellman Certificates

A finite Bellman certificate supplies asymptotic comparison points for undiscounted planning with a state-dependent gain. Algorithm 1is inspired by the approximately shifted Halpern update ofZurek & Chen (2025, Algorithm 1), with the robust operator T. Phase I computes $x_N = T^N 0$ and uses it both to estimate the gain as $\widehat{g}_N = x_N/N$ and to initialize Phase II. The second phase anchors at x_N and shifts Bellman updates by this estimate. We assume T can be exactly computed and applied.

Algorithm 1 Approximately Shifted Robust Halpern Iteration

Input: $N \ge 1$ and an exact robust Bellman oracle T.
Initialize: $x_0 \leftarrow 0$.
`Phase I: Gain estimation and warm start.`
for $j = 0, 1, \ldots, N-1$ **do**
  $x_{j+1} \leftarrow T x_j$.
end for
$\widehat{g}_N \leftarrow x_N/N$ and $z_0 \leftarrow x_N$.
`Phase II: Approximately shifted Halpern iteration.`
for $t = 0, 1, \ldots, N-1$ **do**
  $z_{t+1} \leftarrow \dfrac{2}{t+3} z_0 + \dfrac{t+1}{t+3}\big(T z_t - \widehat{g}_N\big)$.
end for
$Z_N \leftarrow z_N$.
Evaluate the action scores at Z_N to obtain TZ_N and choose
  $\pi_N(i) \in \arg\max_{a \in A(i)} \{ r_{ia} + \min_{p \in \mathcal{U}_{ia}} p^\top Z_N \}$ for every $i \in S$.
Output: $\widehat{g}_N$, Z_N, $TZ_N - Z_N$, and π_N.

Theorem 6.1. *Under the standing finite-state, finite-action, compact post-action (s,a)-rectangular model, suppose equation 9 has a finite solution. Then Algorithm 1 satisfies, in supremum norm,*

$$\widehat{g}_N \longrightarrow g^\star, \qquad TZ_N - Z_N \longrightarrow g^\star, \qquad Z_N/N \longrightarrow g^\star. \tag{14}$$

Moreover, there is a finite N_0 such that every output π_N at any $N \ge N_0$ satisfies $g^{\pi_N} = g^\star$, and each such controller is optimal from every initial state against history-dependent nature.

For a finite solution (g,h), Theorem 4.1 and Lemma I.1 imply the affine defect $\omega_h(t) := \|T(h+tg) - h - (t+1)g\|_\infty \longrightarrow 0$. This relation, however, need not become exact at any finite t: under general compact ambiguity, rows outside the gain face can remain preferable when their bias advantage offsets

a small loss in continuation gain (Appendix L gives an example with $\omega_h(t) > 0$ for every sufficiently large finite t). We therefore retain the vanishing defect $\omega_h(t)$ in the Halpern comparison and establish both $TZ_N - Z_N \to g^\star$ and $Z_N/N \to g^\star$. The first limit controls one-step reward balance; the second eventually excludes each controller action with $m_{ia}(g^\star) < g_i^\star$ from the greedy rule. At every remaining action, every feasible nature row satisfies $p^\top g^\star \geq g_i^\star$, so the displacement error bounds the controller's statewise robust gain loss. There are finitely many deterministic controllers; hence vanishing loss makes every sufficiently late greedy controller exactly optimal from all states. This gain-active identification is the additional step required when the optimal gain is a vector. The affine defect also affects the gain estimator's rate, which is discussed in Appendix L.

Remark 6 (Relation to discounted and constant-gain planning)**.** *Discounted reductions connect robust average reward to discounted planning (Wang et al., 2023d; Grand-Clément & Petrik, 2023; Roch et al., 2025; Yang et al., 2026; Grand-Clement et al., 2023). Discounted-based planning typically directly solves for a discounted robust MDP with large enough discount factor (Roch et al., 2025; Grand-Clément & Petrik, 2023), or with an increasing factor (Wang et al., 2023e; Grand-Clement et al., 2023). The planner studied here uses the undiscounted operator and estimates its vector drift directly. Existing direct anchored value iteration and robust Bellman methods provide direct approaches to average-reward planning (Wang et al., 2023d;e; Roch et al., 2026; Xu et al., 2025a;b) under the constant gain settings, where the translation identity $T(x + c\mathbf{1}) = Tx + c\mathbf{1}$ turns $Th = h + \rho\mathbf{1}$ into a fixed-point equation modulo constant vectors. However, a state-dependent gain retains its drift after this scalar normalization. Ours instead controls the vector displacement $TZ_N - Z_N \to g^\star$ through an estimated shift and asymptotically affine comparison points.*

Numerical verification. We further numerically evaluate Algorithm 1 against a nominal counterpart using the same anchored updates (Zurek & Chen, 2025) on three multichain models: boundary leakage, periodic recurrent classes, and a safe-risky decision. Figure 1 supports the theoretical convergence of Algorithm 1 and shows that its extracted controllers attain the optimal robust gain in all three models. The nominal method converges for its reference model but selects controllers with strictly smaller worst-case gains. Appendix B provides all details and further empirical analysis.

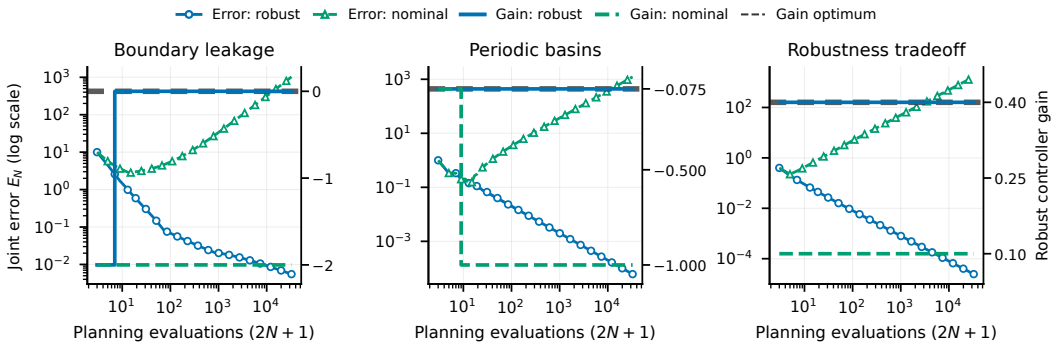

Figure 1: Robust and nominal planning against iterations. Left logarithmic axis: robust-target joint error $E_N = \max\{\|\widehat{g}_N - g^\star\|_\infty, \|Tv_N - v_N - g^\star\|_\infty\}$ (lower is better). Right linear axis: exact robust gain for a specific initial state $g_x^{\pi_N}$ of the extracted controller (higher is better); the black dashed line marks the optimum obtained by solving Bellman equations.

7 Conclusion

We developed a vector Bellman certificate for multichain robust average-reward MDPs. Our coupled gain-bias system preserves gain priority for both players and certifies stationary saddle strategies against history-dependent opponents from every initial state. Its solvability criterion connects stationary gain attainment to uniform control of canonical transient corrections, with concrete sufficient conditions permitting distinct recurrent-class gains. This certificate also provides the asymptotic structure needed for direct undiscounted planning. Under finite solvability, our approximately shifted Halpern iteration recovers the vector gain through its estimates and Bellman displacements, and its greedy controllers are eventually average-optimal. Our studies thus provided comprehensive and systemic understandings of robust average-reward MDPs beyond constant gains.

GENERATIVE AI DISCLOSURE

We used generative AI tools to help review and polish proofs, and identify potential gaps in arguments. Additionally, we used these tools to assist with manuscript organization, language editing, and LaTeX preparation. We did not use 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, design experiments, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

We have reviewed all AI-assisted work. In particular, we checked the modified proof arguments and polished writings, 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 multichain robust MDPs in this paper hold only under the stated modeling assumptions. Any practical deployment should separately consider domain-specific requirements concerning safety, fairness, privacy, security, and legal compliance.

REPRODUCIBILITY STATEMENT

This paper is mainly theoretical with simple illustration experiments without any external datasets or empirical benchmarks. The appendices provide complete proofs of the theory, and we provide all details for reproducing results in appendix.

REFERENCES

Jinane Abounadi, Dimitri P. Bertsekas, and Vivek S Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

Marianne Akian and Stéphane Gaubert. Spectral theorem for convex monotone homogeneous maps, and ergodic control. *Nonlinear Analysis: Theory, Methods & Applications*, 52(2):637–679, 2003.

Marianne Akian, Jean Cochet-Terrasson, Sylvie Detournay, and Stéphane Gaubert. Policy iteration algorithm for zero-sum multichain stochastic games with mean payoff and perfect information. *arXiv preprint arXiv:1208.0446*, 2012.

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *Proc. International Conference on Machine Learning (ICML)*, pp. 511–520. PMLR, 2021.

Amitabh Basu, Kipp Martin, and Christopher Thomas Ryan. Projection: a unified approach to semi-infinite linear programs and duality in convex programming. *arxiv preprint arXiv:1304.3030v2*, 2013.

Truman Bewley and Elon Kohlberg. The asymptotic theory of stochastic games. *Mathematics of Operations Research*, 1(3):197–208, 1976. doi: 10.1287/moor.1.3.197. URL `https://pubsonline.informs.org/doi/10.1287/moor.1.3.197`.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, pp. 719–726, 1962.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Jérôme Bolte, Stéphane Gaubert, and Guillaume Vigeral. Definable zero-sum stochastic games. *Mathematics of Operations Research*, 40(1):171–191, 2015.

Krishnendu Chatterjee, Ehsan Kafshdar Goharshady, Mehrdad Karrabi, Petr Novotnỳ, and Đorđe Žikelić. Solving long-run average reward robust mdps via stochastic games. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, 2024.

Zijun Chen, Shengbo Wang, and Nian Si. Sample complexity of distributionally robust average-reward reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 85402–85463, 2025.

E. V. Denardo and B. L. Fox. Multichain Markov renewal programs. *SIAM Journal on Applied Mathematics*, 16(3):468–487, 1968. doi: 10.1137/0116038. URL https://epubs.siam.org/doi/10.1137/0116038.

Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized MDPs and the equivalence between robustness and regularization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Larry G. Epstein and Martin Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31, 2003. doi: 10.1016/S0022-0531(03)00097-8. URL https://people.bu.edu/lepstein/files-research/RMP-JET2003.pdf.

Stéphane Gaubert and Jeremy Gunawardena. The perron-frobenius theorem for homogeneous, monotone functions. *Transactions of the American Mathematical Society*, 356(12):4931–4950, 2004.

Debamita Ghosh, George K. Atia, and Yue Wang. ORVIT: Near-optimal online distributionally robust reinforcement learning. In *Annual AAAI Conference on Artificial Intelligence*, volume 40, pp. 21278–21286, 2026a. doi: 10.1609/aaai.v40i25.39273.

Debamita Ghosh, George K. Atia, and Yue Wang. Online robust reinforcement learning with general function approximation. In *Proc. International Conference on Machine Learning (ICML)*, 2026b.

Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

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-Clement, Marek Petrik, and Nicolas Vieille. Beyond discounted returns: Robust markov decision processes with average and blackwell optimality. *arXiv preprint arXiv:2312.03618v3*, 2023.

Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Proc. International Conference on Machine Learning (ICML)*, 2025.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast Bellman updates for robust MDPs. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1979–1988. PMLR, 2018.

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.

A. Hordijk and L. C. M. Kallenberg. Linear programming and Markov decision chains. *Management Science*, 25(4):352–362, 1979. doi: 10.1287/mnsc.25.4.352. URL https://pubsonline.informs.org/doi/10.1287/mnsc.25.4.352.

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.

Elon Kohlberg. Invariant half-lines of nonexpansive piecewise-linear transformations. *Mathematics of Operations Research*, 5(3):366–372, 1980.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Y Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 59477–59501, 2023a.

Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. An efficient solution to s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13642*, 2023b.

Jongmin Lee and Ernest Ryu. Optimal non-asymptotic rates of value iteration for average-reward markov decision processes. In *Proc. International Conference on Learning Representations (ICLR)*, volume 2025, pp. 32823–32865, 2025.

Yan Li and Alexander Shapiro. Rectangularity and duality of distributionally robust Markov decision processes. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02297-y. URL https://link.springer.com/article/10.1007/s10107-025-02297-y.

Zhipeng Liang, Xiaoteng Ma, Jose Blanchet, Jiheng Zhang, and Zhengyuan Zhou. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.

Felix Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15:405–418, 2021. doi: 10.1007/s11590-020-01617-9. URL https://link.springer.com/article/10.1007/s11590-020-01617-9.

Thomas M. Liggett and Steven A. Lippman. Stochastic games with perfect information and time average payoff. *SIAM Review*, 11(4):604–607, 1969. doi: 10.1137/1011093. URL https://epubs.siam.org/doi/10.1137/1011093.

Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust Markov decision processes. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 701–709, 2013.

Zhishuai Liu and Pan Xu. Minimax optimal and computationally efficient algorithms for distributionally robust offline reinforcement learning. *arXiv preprint arXiv:2403.09621*, 2024.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 13623–13643. PMLR, 2022.

Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms. *arXiv preprint arXiv:2404.03578*, 2024.

Xiaoteng Ma, Xiaohang Tang, Li Xia, Jun Yang, and Qianchuan Zhao. Average-reward reinforcement learning with trust region methods. *arXiv preprint arXiv:2106.03442*, 2021.

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1):159–195, 1996a.

Sridhar Mahadevan. An Average-Reward Reinforcement Learning Algorithm for Computing Bias-Optimal Policies. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1*, AAAI'96, pp. 875–880. AAAI Press, 1996b. ISBN 0-262-51091-X. event-place: Portland, Oregon.

Jean-François Mertens and Abraham Neyman. Stochastic games. *International Journal of Game Theory*, 10:53–66, 1981. doi: 10.1007/BF01769259. URL https://link.springer.com/article/10.1007/BF01769259.

Arnab Nilim and Laurent El Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 839–846, 2004.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 9582–9602. PMLR, 2022.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.

Kishan Panaganti, Adam Wierman, and Eric Mazumdar. Model-free robust ϕ-divergence reinforcement learning using both offline and online data. *arXiv preprint arXiv:2405.05468*, 2024.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1 edition, April 2014. ISBN 978-0-471-61977-2 978-0-470-31688-7.

Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 100–108. PMLR, 2024.

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.

Andrzej Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125:235–261, 2010. doi: 10.1007/s10107-010-0393-3. URL `https://link.springer.com/article/10.1007/s10107-010-0393-3`. See also the erratum, Mathematical Programming 145:601–604 (2014), doi:10.1007/s10107-014-0783-z.

Paul J Schweitzer. On undiscounted markovian decision processes with compact action spaces. *RAIRO-Operations Research-Recherche Opérationnelle*, 19(1):71–86, 1985.

Paul J Schweitzer and Awi Federgruen. The functional equations of undiscounted markov renewal programming. *Mathematics of Operations Research*, 3(4):308–321, 1978.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

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. *arXiv preprint arXiv:2305.16589*, 2023.

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.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust MDPs using function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 181–189. PMLR, 2014.

Yi Wan and Richard S Sutton. On convergence of average-reward off-policy control algorithms in weakly communicating mdps. *arXiv preprint arXiv:2209.15141*, 2022.

Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward markov decision processes. In *Proc. International Conference on Machine Learning (ICML)*, pp. 10653–10662. PMLR, 2021.

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

He Wang, Laixi Shi, and Yuejie Chi. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024a.

Kaixin Wang, Uri Gadot, Navdeep Kumar, Kfir Levy, and Shie Mannor. Bring your own (non-robust) algorithm to solve robust mdps by estimating the worst kernel. *arXiv preprint arxiv: 2306.05859*, pp. arXiv–2306, 2023a.

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:2509.14203v3*, 2025.

Shengbo Wang and Nian Si. Non-rectangular average-reward robust mdps: Optimal policies and their transient values. *arXiv preprint arXiv:2603.00945*, 2026.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3370–3398. PMLR, 2023b.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust q-learning. *arXiv preprint arXiv:2305.18420*, 2023c.

Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024b.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 7193–7206, 2021.

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 162, pp. 23484–23526. PMLR, 2022.

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, 2023d.

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, 2023e.

Yue Wang, Zhongchang Sun, and Shaofeng Zou. A unified principle of pessimism for offline reinforcement learning under model mismatch. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024c.

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

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.

Shangtong Zhang, Yi Wan, Richard S Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 12578–12588. PMLR, 2021.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3331–3339. PMLR, 2021.

Bruno Ziliotto. A tauberian theorem for nonexpansive operators and applications to zero-sum stochastic games. *Mathematics of Operations Research*, 41(4):1522–1534, 2016.

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)*, 2024.

Matthew Zurek and Yudong Chen. Faster fixed-point methods for multichain mdps. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Appendix organization. Appendices A and B discuss related work and give the details of the experiments. Appendices C–H establish the strategic values and stationary guarantees without assuming a finite bias. Appendix I verifies the vector Bellman system, and Appendix J characterizes its solvability. Appendix K studies the structure and geometry of compatible biases. Appendix L uses the resulting asymptotically affine trajectory to prove convergence of the planning algorithm. Appendix M collects sufficient conditions for solvability, obstruction examples, and invariance properties.

A Related work

The paper connects robust dynamic programming, multichain average-reward theory, and the analysis of nonexpansive operators. The central distinction is between the existence of a long-run strategic value and its representation by a finite gain-bias pair. We organize the literature around this distinction and its consequences for planning.

A.1 Transition ambiguity, rectangularity, and dynamic consistency

Robust and distributionally robust MDPs. Robust MDPs optimize a policy against a family of plausible transition models. Classical robust dynamic programming identifies rectangularity assumptions under which local worst-case transition choices yield a recursive description of the value (Iyengar, 2005; Nilim & El Ghaoui, 2004; Wiesemann et al., 2013). Distributionally robust formulations also model uncertainty through distributions over model parameters and allow statistical information to enter the ambiguity description (Xu & Mannor, 2010). The precise uncertainty object and the information available to nature matter: an uncertainty set over transition rows, a distribution over kernels, and a single unknown kernel chosen at the outset need not define the same control problem. Here nature selects a transition row after observing the current state and action, with independent admissibility constraints across state-action pairs. This post-action (s, a)-rectangular structure determines the order of optimization in our Bellman operator.

Dynamic consistency and risk-averse control. Rectangularity has a broader interpretation in sequential decision theory. Epstein & Schneider (2003) connect rectangular sets of priors to recursive multiple-priors preferences and dynamic consistency. In Markov control, Ruszczyński (2010) develop dynamic programming with Markov risk measures for finite-horizon and discounted problems. The connection to our operator is visible at the one-step level: the lower expectation $\min_{p\in\mathcal{U}_{sa}} p^\top v$ equals the negative of the upper expectation $\max_{p\in\mathcal{U}_{sa}} p^\top(-v)$. Thus worst-case reward evaluation has a natural risk-averse interpretation after reversing signs. These connections explain the recursive structure of robust evaluation; the existence of a finite bias for an undiscounted, state-dependent gain requires additional long-run analysis.

Coupled uncertainty and the timing of nature's choices. Rectangularity can also be imposed on a representation of uncertainty rather than directly on individual transition rows. Goyal & Grand-Clement (2023) study factor-matrix uncertainty that couples transitions across states while retaining tractability under rectangularity in the factor representation. Li & Shapiro (2025) examine the relationship between static and game formulations of distributionally robust MDPs and the role of rectangularity in their equivalence and duality. For average reward, Wang & Si (2026) study nonrectangular uncertainty with a stationary kernel chosen by nature and history-dependent controller policies. These models address different forms of dependence and information. Our results concern the stagewise post-action model; this specification is essential to the gain-restricted row sets and the controller-nature comparisons used below.

A.2 Multichain average reward and stochastic-game values

Classical multichain optimality equations. Average-reward MDPs model continuing decisions without an exogenous discount factor (Puterman, 2014). The distinction between long-run gain and transient bias is classical. Blackwell (1962) establish the connection between discounted optimization near discount factor one and undiscounted optimality in finite models. Denardo & Fox (1968) develop multichain Markov renewal programming, while Schweitzer & Federgruen (1978) study the solution structure of undiscounted functional equations, including the degrees of freedom associated with optimal recurrent behavior. In general multichain models the gain can depend on the initial state.

Gain-bias equations then compare continuation gains first and rewards and biases among gain-optimal actions second. Under a fixed policy and kernel, recurrent-class rewards and absorption probabilities determine the gain vector. Transition ambiguity adds a second optimization that can change those classes and probabilities. Our vector Bellman system uses the same gain-first principle while requiring compatible comparisons for both players.

Finite stochastic games and mean payoff. The long-run value problem also belongs to the theory of zero-sum stochastic games. For finite state and action spaces, Bewley & Kohlberg (1976) establish a common asymptotic limit of normalized finite-horizon and discounted values, and Mertens & Neyman (1981) prove existence of the uniform value. The latter is a strategic guarantee across sufficiently long horizons; it does not in general imply that both players have stationary optimal strategies. Perfect-information games have additional structure, with classical stationary-strategy results for time-average payoff (Liggett & Lippman, 1969). Their multichain gain-bias structure and policy iteration are developed further by Akian et al. (2012). These are direct precedents for robust models with finitely many effective nature actions.

For polytopic (s, a)-rectangular ambiguity, minimizing a linear continuation value can be reduced to the finitely many extreme rows. Chatterjee et al. (2024) exploit the resulting connection to finite turn-based stochastic games to obtain long-run robust planning and complexity results. Consequently, direct average-reward planning beyond scalar-gain assumptions already has precedents in the polytopic case. Our analysis also permits compact curved row sets, where a finite reduction need not be available and finite Bellman solvability must be examined separately.

Compact-action games and asymptotic values. Finiteness of the state space alone does not replace assumptions on the action sets in general stochastic-game value theory. Bolte et al. (2015) use definability and additional structural conditions to establish uniform values for classes of compact-action games, including definable perfect-information games. Definability includes semialgebraic examples and supplies regularity beyond compactness. A complementary operator approach relates convergence of normalized discounted and finite-horizon values through Tauberian theorems (Ziliotto, 2016). We use the latter connection after establishing discounted convergence for the present robust model. The relevant conclusion is convergence of normalized values for arbitrary compact row sets, without imposing definability. It should be distinguished from both a finite gain-bias representation and stronger uniform-strategy conclusions for general compact-action games.

Robust average-reward theory. Robust average-reward Bellman equations and algorithms have been developed under unichain assumptions on the policy-kernel family (Wang et al., 2023d; 2024d). More recent theory gives conditions for scalar robust Bellman solvability under broader communication and information structures, including one-sided weak communication (Wang & Si, 2025). Such assumptions can allow multiple recurrent classes for some choices while still producing an optimal gain independent of the initial state. Thus the relevant distinction for this paper is state dependence of the optimal gain, rather than simply whether any multichain transition matrix is admissible.

For compact (s, a)-rectangular sets, Grand-Clement et al. (2023) establish deterministic stationary controller optimality, strong duality, equivalence of the principal average-payoff conventions, and normalized discounted convergence without a unichain assumption. They also show that nature's stationary worst case need not be attained. These results provide the strategic foundation for our study. We formulate the all-state guarantees needed by the Bellman analysis and obtain normalized finite-horizon convergence using the Tauberian connection. The subsequent questions are whether the value admits a finite vector gain-bias certificate, which stationary choices such a certificate supports, and how to recover the gain and a controller policy by direct iteration.

A.3 Finite biases, nonlinear operators, and feasibility geometry

Compact-action Bellman solvability. The fixed-policy criterion comes from classical compact-action MDP theory. After reversing the reward sign, Schweitzer (1985, Theorem 1) characterizes finite one-player gain-bias solvability through stationary gain attainment and a uniform lower bound on canonically normalized biases. The bound controls transient corrections as transition kernels vary; pointwise finiteness for each kernel is insufficient. We apply this criterion to the nature problem induced by a fixed controller policy. For robust optimal control, the proof combines a lower barrier obtained from this one-player result with an upper barrier supplied by a full nature plan. The

two-player formulation makes explicit which policies, replies, and common bias bounds must be compatible.

Nonlinear spectral theory and recurrent classes. Undiscounted Bellman operators are monotone, additively homogeneous with respect to scalar constants, and nonexpansive in the sup norm. Nonlinear Perron-Frobenius theory provides bounded-orbit criteria for additive eigenvectors and fixed points (Gaubert & Gunawardena, 2004). For convex monotone homogeneous maps, critical classes describe the structure and degrees of freedom of eigenspaces (Akian & Gaubert, 2003). These results connect recurrent behavior to Bellman solvability. A robust max-min operator, however, need not be convex, so the convex spectral theorem does not apply to it directly. Our analysis instead fixes controller-nature selector pairs, uses the classical Poisson representation for their induced chains, and imposes both players' deviation inequalities on the remaining recurrent-class offsets. This identifies when one finite bias supports all required comparisons.

Linear programming and semi-infinite certificates. Linear programming is another classical route to average-reward control. Hordijk & Kallenberg (1979) formulate finite MDP average optimization through a single linear program and relate its feasible solutions to stationary policies. In the present compact-row model, requiring a bias inequality for every admissible deviation produces a semi-infinite feasibility problem: there are finitely many bias coordinates but potentially infinitely many constraints. Projection and duality methods for such systems are developed by Basu et al. (2013). Our mixed constraint cone combines controller and nature deviations. Its closure records limiting inconsistencies that can arise even when no finite combination gives an exact contradiction, and a reward-to-flow ratio determines the minimum compatible bias span. The separation principles are standard; their role here is to characterize a common two-player Bellman certificate and quantify its size.

A.4 Direct planning and anchored iterations

Invariant half-lines and approximate affine behavior. For finite-action perfect-information games and polytopic robust MDPs, the Bellman operator is piecewise affine. Kohlberg (1980) show that a nonexpansive piecewise-linear map admits an invariant half-line, which describes an eventual affine trajectory with a fixed growth direction. This structure underlies multichain game algorithms (Akian et al., 2012). For general compact row sets, a finite Bellman solution can instead yield an asymptotically affine trajectory: its one-step defect tends to zero, but the trajectory need not become exactly invariant after a finite threshold. The distinction matters for planning because an argument based on eventual exact equality does not automatically cover curved ambiguity sets. Our convergence analysis tracks this vanishing defect explicitly.

Halpern iteration and multichain planning. Anchored fixed-point methods originate in the iteration of Halpern (1967). Quantitative analyses include sharp residual bounds for nonexpansive maps in Hilbert spaces (Lieder, 2021). Those results explain the general anchoring mechanism, while Bellman planning requires estimates in the operator's relevant norm and may involve a nonzero growth direction rather than an ordinary fixed point. Recent nominal MDP work develops anchored and shifted methods that address these issues (Lee & Ryu, 2025; Zurek & Chen, 2025). In particular, our update follows the approximately shifted iteration of Zurek & Chen (2025). The robust analysis controls the additional affine defect and connects gain and displacement estimates to controller-policy extraction. The iteration's origin, its robust convergence argument, and the structural conditions ensuring a finite bias are therefore separate parts of the comparison.

Computing robust Bellman updates. Iteration complexity and the cost of each inner minimization are complementary questions. Ho et al. (2018) develop efficient exact Bellman updates for ℓ_1 ambiguity, and Ho et al. (2021) combine efficient updates with partial policy iteration for discounted robust MDPs. Such methods can serve as computational components when the row sets in our model have the corresponding structure. They do not by themselves provide an undiscounted multichain convergence argument. Conversely, an operator-level convergence result for arbitrary compact row sets does not imply a uniformly efficient implementation of every inner optimization problem; its computational use depends on how the ambiguity sets are represented.

A.5 AVERAGE-REWARD AND ROBUST REINFORCEMENT LEARNING

Nominal average-reward learning. Average-reward reinforcement learning includes differential value estimation, temporal-difference methods, and Q-learning (Mahadevan, 1996a;b; Abounadi et al., 2001; Zhang et al., 2021; Ma et al., 2021). Regret-based work such as Jaksch et al. (2010) studies exploration in unknown communicating MDPs using a diameter parameter. Other analyses develop convergence under weak communication and finite-sample guarantees governed by bias span or related structural quantities (Wan et al., 2021; Wan & Sutton, 2022; Wan et al., 2024; Zurek & Chen, 2024). These results show why communication, recurrent structure, and bias size are central to both learning and planning. Our setting isolates the deterministic robust planning and solvability questions with access to the Bellman operator; a statistical learning guarantee would additionally need to control how transition-estimation errors affect those quantities.

Robust average-reward learning. Existing methods include relative-value TD and Q-learning, policy optimization, discounted reductions, anchored procedures, and stochastic approximation (Wang et al., 2023e; Sun et al., 2024; Roch et al., 2025; Chen et al., 2025; Xu et al., 2025b;a; Roch et al., 2026; Yang et al., 2026; Wang et al., 2025). Their assumptions vary, with many guarantees using unichain, irreducibility, or uniform ergodicity conditions that yield a state-independent robust gain. These conditions provide ways to control long-run sensitivity and transient behavior. Our finite-bias characterization addresses the structural question that arises when the gain can vary across initial states: which robust models still admit a finite certificate on which a direct planning analysis can be based?

Discounted and finite-horizon robust learning. A large literature studies statistical estimation of worst-case values under discounted or finite-horizon criteria. Early sample-based approaches include robust temporal-difference learning, approximate dynamic programming, and linear policy evaluation (Lim et al., 2013; Tamar et al., 2014; Badrinath & Kalathil, 2021; Wang & Zou, 2021). Subsequent tabular analyses cover model-based estimation and plug-in planning, as well as model-free procedures, for several ambiguity families and data-access models (Yang et al., 2022; Panaganti & Kalathil, 2022; Xu et al., 2023; Shi et al., 2023; Zhou et al., 2021; Wang et al., 2023b; Liang et al., 2023; Liu et al., 2022; Wang et al., 2023c; 2024b; 2023a; Kumar et al., 2023b; Derman et al., 2021). Online robust learning additionally treats exploration (Wang & Zou, 2021; Lu et al., 2024; Ghosh et al., 2026a; He et al., 2025). Related model-free methods use robust Q-learning, multilevel Monte Carlo, or variance reduction (Liu et al., 2022; Wang et al., 2023b; 2024b;a; Ghosh et al., 2026b), while policy-based approaches analyze robust policy-gradient and actor-critic methods (Wang & Zou, 2022; Kumar et al., 2023a).

Offline learning and function approximation. Robust offline methods combine ambiguity-aware pessimism with tabular or fitted value iteration and structured function approximation (Zhou et al., 2021; Panaganti et al., 2022; Shi & Chi, 2022; Blanchet et al., 2023; Liu & Xu, 2024; Panaganti et al., 2024; Wang et al., 2024a;c). Other work considers distributional robustness with function approximation or simulator access in large or continuous state spaces (Ramesh et al., 2024). These literatures address estimation, coverage, and approximation errors. Their discounted contraction or finite-horizon recursion controls the propagation of those errors. In the undiscounted multichain problem studied here, finite-bias solvability and state-dependent growth must first be understood to obtain an analogous foundation for algorithmic analysis.

B NUMERICAL EXPERIMENTS

We examine convergence of the computed gain and Bellman displacement, and robust average-reward performance of the extracted controller. The three models are designed stress tests in which midpoint nominal parameters favor an action with a smaller worst-case gain. The first two have curved, nonpolytopic ambiguity and isolate boundary leakage and periodicity, respectively. The third is a polytopic safe-risky decision. Each model has a finite gain-bias certificate, specified below, so the solvability hypothesis of Theorem 6.1 is satisfied.

B.1 METHODS, BUDGETS, AND EVALUATION

Methods and initialization. Let T denote the robust Bellman operator and let T_0 replace each uncertain row set by the nominal reference row specified below. We compare Algorithm 1, using

T, with a nominal ablation using T_0. The latter is the approximately shifted multichain iteration of Zurek & Chen (2025), applied to the fixed nominal model. For either planning operator $\mathcal{T} \in \{T, T_0\}$, the run starts at zero, forms $x_N := \mathcal{T}^N 0$, sets $\widehat{g}_N := x_N/N$ and $z_0 := x_N$, and performs

$$z_{t+1} := \frac{2z_0 + (t+1)(\mathcal{T} z_t - \widehat{g}_N)}{t+3}, \qquad t = 0, \dots, N-1.$$

The output vector is $v_N := z_N$ (denoted Z_N in the planning theorem). One final evaluation of $\mathcal{T} v_N$ extracts a controller greedy for that method's own operator. Both methods therefore use $2N+1$ planning evaluations. Tied action values are resolved in favor of c in Models I and II and the risky action in Model III. The nominal controller is evaluated under the robust model with its selected action fixed.

Budget convention. We evaluate every integer $N = 1, \dots, 16384$, including both parities. Each point represents a complete run from zero with that budget; the anchor and estimated gain depend on N. Thus $2N+1$ measures the planning budget of a run. The implementation shares common first-phase iterates and batches the independent second-phase calculations, producing the same outputs as individually initialized runs. Batched execution time is recorded separately from the per-run oracle count. The nominal method uses one additional robust operator evaluation to measure its robust displacement. This external diagnostic is excluded from its nominal planning budget and is never fed into its updates or policy selection.

Convergence errors. All methods are evaluated against the exact robust gain $g^\star$. Define

$$E_g(N) := \|\widehat{g}_N - g^\star\|_\infty, \qquad E_d(N) := \|T v_N - v_N - g^\star\|_\infty, \qquad E_N := \max\{E_g(N), E_d(N)\}. \tag{15}$$

The main figure uses E_N to show both convergence targets in one panel per model. The top and middle rows of Figure 2 display the two components separately. For Algorithm 1, Theorem 6.1 gives $E_N \to 0$ under finite Bellman solvability. This is an asymptotic statement and does not require the errors to decrease at every finite budget. For the nominal method, this robust-target error includes disagreement between the nominal and robust objectives. We separately measure each solver's own-objective error:

$$E_N^{\text{own}} := \max\{\|\widehat{g}_N - g^\star_{\mathcal{T}}\|_\infty, \|\mathcal{T} v_N - v_N - g^\star_{\mathcal{T}}\|_\infty\},$$

where $g^\star_{\mathcal{T}} := g^\star$ for $\mathcal{T} = T$ and $g^\star_{\mathcal{T}} := g^\star_0$ for $\mathcal{T} = T_0$, with $g^\star_0$ the optimal gain of the nominal reference model. The bottom row of Figure 2 reports this diagnostic. The data retain full output vectors and statewise robust policy gains.

Robust evaluation of the output controller. Each model has one decision state x. For the controller π_N selected by each method, we plot its actual worst-case average reward

$$G_N := g_x^{\pi_N} = \inf_{q \in \mathcal{Q}_S} \eta_x^{\pi_N, q}, \qquad G^\star := g_x^\star. \tag{16}$$

This is evaluated analytically from the selected controller, rather than estimated by its value iterate or by a finite simulated rollout. In the first two models, state y has the same gain as x, and all remaining states have policy-independent gains. The latter property also holds in the third model. Consequently, in every experiment,

$$\|g^\star - g^{\pi_N}\|_\infty = G^\star - G_N.$$

Reaching the horizontal reference $G^\star$ therefore verifies optimality from every initial state for these models. Policy gains are plotted on linear axes, including negative values.

Exact row minimization and implementation. Linear minimization over the convex hull of a curve has the same value as minimization over the generating curve. In Model I, the row objective is $v_x + u(v_y - v_x) + u^3(v_z - v_x)$; we compare both endpoints and any real stationary point in $[0, 1/2]$. In Model II it is quadratic, so both endpoints and any interior minimizing vertex suffice. The alternative-action row objectives and the Model III objective are affine, so their minima occur at interval endpoints. All calculations use these analytic minimizers in double precision, with no discretization of the ambiguity set and no Monte Carlo policy evaluation. The computations are deterministic. The nominal reference rows are specified as part of each model.

B.2 MODEL I: BOUNDARY LEAKAGE

The ordered state space is (x, y, z, w, H). State x has actions c (curved) and t (risky); every other state has one action. The rewards are $r(x, c) := 0$, $r(x, t) := 10$, $r(y) := -1$, $r(z) := 1$, $r(w) := -2$, and $r(H) := 5$. The uncertain rows are

$$\mathcal{U}_{x,c} := \mathrm{co}\{(1 - u - u^3, u, u^3, 0, 0) : 0 \le u \le 1/2\},$$
$$\mathcal{U}_{x,t} := \{(0, 0, 0, 1 - \rho, \rho) : 0 \le \rho \le 1\}.$$

State y returns to x, and z, w, H are absorbing. These remaining rows are known. The nominal parameters are the interval midpoints $u_0 := 1/4$ and $\rho_0 := 1/2$.

The exact robust gain and a gain-face bias are

$$g^\star = (0, 0, 1, -2, 5), \qquad h := (0, -1, 0, 0, 0).$$

The curved continuation gain is u^3, minimized at $u = 0$, whereas the risky continuation gain is $-2 + 7\rho$, minimized at $\rho = 0$. Thus only c is gain-active, and its minimizing gain face is the self-loop. The stated bias satisfies the restricted equation there and the deterministic equations at the other states. The curved controller has robust gain $g^\star$, while the risky controller has gain $(-2, -2, 1, -2, 5)$. Hence $G_N \in \{0, -2\}$.

Under the nominal model, action c eventually reaches z and has gain 1 at x, y. Action t has nominal gain $(1 - \rho_0)(-2) + 5\rho_0 = 3/2$, so it is strictly preferable nominally. The nominal optimal gain is $g_0^\star = (3/2, 3/2, 1, -2, 5)$, and its optimal controller has robust loss 2. The immediate reward 10 affects transient decisions but contributes zero to the long-run average after absorption.

The curved rows create long transients near $u = 0$: every fixed $u > 0$ eventually leads to z, while $u = 0$ leaves x recurrent. This change in recurrent structure makes the model a test of planning near the boundary of the uncertainty set.

B.3 MODEL II: PERIODIC BASINS

The ordered states are $(x, y, h_0, h_1, m, \ell)$. State x has actions c (curved) and f (risky), both with reward zero. Every other state has one action. The rewards at (y, h_0, h_1, m, ℓ) are $(-0.3, 1, -1, -1, 1)$. For the curved action,

$$\mathcal{U}_{x,c} := \mathrm{co}\{p(u) : 0 \le u \le 1\},$$

where, in the stated coordinate order,

$$p(u) := (0, 0, 0.55 + 0.10u, 0, 0.25 - 0.05u^2, 0.20 - 0.10u + 0.05u^2).$$

The risky row set is

$$\mathcal{U}_{x,f} := \{(0, 0, 0, 0, 1 - \rho, \rho) : 0 \le \rho \le 1\}.$$

State y goes to x, and h_0, h_1 alternate deterministically. States m, ℓ are absorbing. The nominal parameters are $u_0 := 1/2$ and $\rho_0 := 1/2$.

The exact robust certificate is

$$g^\star = (-3/40, -3/40, 0, 0, -1, 1), \qquad h := (27/40, 9/20, 1, 0, 0, 0).$$

Indeed, $p(u)^\top g^\star = -3/40 + (u - 1/2)^2/10$, uniquely minimized at $u = 1/2$, and $p(1/2)^\top h = 3/5 = g_x^\star + h_x$. The risky continuation gain is $2\rho - 1$, whose worst value is -1. Thus c is strictly gain-preferred, and the remaining bias equations follow from the deterministic transitions. The curved and risky controllers have robust gains $g^\star$ and $(-1, -1, 0, 0, -1, 1)$, respectively. Consequently, $G_N \in \{-3/40, -1\}$.

Nominally, the risky action has gain $2\rho_0 - 1 = 0$, which exceeds the curved action's $-3/40$. The nominal optimal gain is $g_0^\star = (0, 0, 0, 0, -1, 1)$, and the nominal optimal controller has robust loss $37/40$. The deterministic two-cycle has alternating rewards and zero average gain. This model tests convergence with periodic recurrent dynamics and different gains across recurrent classes.

B.4 Model III: a safe-risky decision

This model gives a direct safe-risky comparison. There are three states (x, L, H). States L, H are absorbing with rewards $0, 1$, respectively. At x, both available actions have reward zero. The safe action has known row $(0, 0.6, 0.4)$, while the risky action has row set

$$\mathcal{U}_{x,\text{risky}} := \{(0, 1-p, p) : 0.1 \le p \le 0.9\}.$$

The nominal risky row uses $p_0 := 0.5$; the safe row is unchanged. The robust gain is $g^\star = (0.4, 0, 1)$, achieved by the safe action, with bias $h := (-0.4, 0, 0)$. These vectors directly satisfy the gain-first system; the row sets are also polytopes. The risky controller has robust gain $(0.1, 0, 1)$, so its robust loss is 0.3. In contrast, nominal planning prefers the risky action and has optimal nominal gain $g_0^\star = (0.5, 0, 1)$. Thus the nominal objective changes the selected controller in this model. This polytopic example complements the two nonpolytopic constructions by making the robustness distinction explicit.

B.5 Results and interpretation

Figure 1 pairs each model's joint error with robust controller gain, and the top and middle rows of Figure 2 separate the gain and displacement errors. Algorithm 1's errors are consistent with the predicted convergence to zero, and its controllers attain the optimal robust gain in all three models. In Model I it selects t at $N = 1, 2$ and c at every tested $N \ge 3$. It selects an optimal controller at every tested budget in Models II and III.

The nominal baseline selects the risky controller at every tested budget in Models I and III. In Model II it selects c at $N = 1, 2, 3$ and f at every tested $N \ge 4$. Its final robust gains are therefore -2, -1, and 0.1, compared with the optimal gains 0, $-3/40$, and 0.4. These gaps follow from the different objectives: the nominal controller optimizes its reference model, whose favorable outcomes are less reliable under worst-case transition evaluation.

At $N = 16384$, Algorithm 1 has joint errors approximately 5.51×10^{-3}, 6.10×10^{-5}, and 2.44×10^{-5} in Models I-III, respectively. The bottom row of Figure 2 shows both methods approaching zero error for their own objectives. The nominal method's robust displacement can nevertheless grow because its output follows the nominal gain direction. Its robust policy losses are 2, $37/40$, and 0.3, respectively. Thus convergence for the reference model and robust controller performance are distinct properties.

C Preliminaries and applicability of prior results

We use the model and notation of Section 2. Throughout, $n := |S|$, $R := \max_{i,a} |r_{ia}|$, $\|\cdot\|_\infty$ is the supremum norm, and $\operatorname{sp}(x) := \max_i x_i - \min_i x_i$. Vector inequalities, maxima, and minima are coordinatewise. A coordinatewise infimum over selectors need not be attained by one selector; whenever a single selector works for every state, we prove this separately. For $x \in \mathbb{R}^S$ and a state-action pair (i, a) we write, as in the main text,

$$m_{ia}(x) := \min_{p \in \mathcal{U}_{ia}} p^\top x.$$

Compactness of $\mathcal{U}_{ia}$ attains this minimum, and $|m_{ia}(x) - m_{ia}(y)| \le \|x - y\|_\infty$ because every row is a probability vector. For a full selector $q \in \mathcal{Q}_S$, the nominal MDP obtained by fixing q has the Bellman operator

$$(H_q x)_i := \max_{a \in A(i)} \{r_{ia} + q_{ia}^\top x\}.$$

Proof dependencies. We use the normalized discounted limit and the approximately optimal stationary kernel for a fixed controller from Grand-Clement et al. (2023, Lemma 4.7 and Theorem 4.3), and the theorem relating discounted and finite-horizon limits from Ziliotto (2016, Theorem 1.2). We verify their hypotheses below. We also use classical finite-chain facts (Puterman, 2014, Chapters 8–9). Discounted dynamic programming, the pathwise comparisons, and weighted stationary duality are proved here. In the duality proof, the stationary saddle needed for a finite restriction follows directly by taking a limit of discounted saddles. This is a special case of the classical perfect-information game result of Liggett & Lippman (1969). All these arguments allow compact nonconvex row

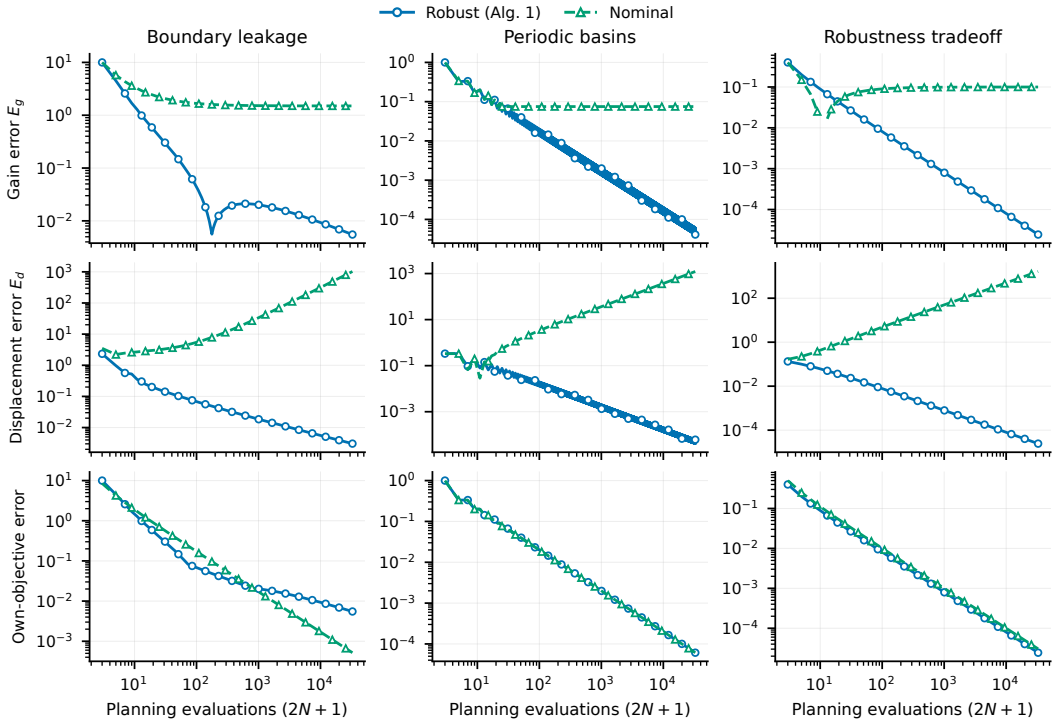

Figure 2: Additional convergence diagnostics at every integer budget. Columns correspond to Models I-III. Top: robust gain-estimation error $E_g(N)$. Middle: robust displacement error $E_d(N)$. Bottom: joint error E_N^{own} against each method's own operator and optimal gain. Both methods converge for their own planning objectives, while the nominal controllers retain the robust policy losses shown in Figure 1.

sets and randomized strategies. The theorem numbers for Grand-Clement et al. (2023) refer to arXiv:2312.03618v3, dated January 14, 2025.

C.1 Rectangular dynamic programming and one-step comparisons

Information structure. Fix a strategy pair. Let $\mathcal{F}_t$ be the σ-field generated by the history

$$(S_0, A_0, p_0, \ldots, S_{t-1}, A_{t-1}, p_{t-1}, S_t)$$

and by the randomization both players realized before stage t; the lotteries of stage t itself are excluded. Let $\mathcal{G}_t$ additionally contain the action A_t, nature's row $p_t \in \mathcal{U}_{S_t A_t}$, and the stage-$t$ lotteries that produced them, but not S_{t+1}. Then $\mathcal{F}_t \subseteq \mathcal{G}_t \subseteq \mathcal{F}_{t+1}$, and the transition rule reads

$$\mathbb{E}[f(S_{t+1}) \mid \mathcal{G}_t] = p_t^\top f \qquad (f \in \mathbb{R}^S). \tag{17}$$

The fresh random seeds used to implement a stage's lotteries are independent of the past. The action and row distributions themselves may depend on the observed history. In particular, if the controller uses $\pi \in \Pi_S$, then $\mathbb{P}(A_t = a \mid \mathcal{F}_t) = \pi(a \mid S_t)$. The enlarged σ-fields are used only to condition expectations in the proof. Each player's strategy still uses only that player's permitted observations.

One-step comparisons. For every $v \in \mathbb{R}^S$ and every $t \geq 0$,

$$\mathbb{E}\big[r_{S_t A_t} + v(S_{t+1}) \;\big|\; \mathcal{F}_t\big] \geq (T^\pi v)(S_t) \qquad \text{if the controller uses } \pi \in \Pi_S, \tag{18}$$

$$\mathbb{E}\big[r_{S_t A_t} + v(S_{t+1}) \;\big|\; \mathcal{G}_t\big] = r_{S_t A_t} + q_{S_t A_t}^\top v \leq (H_q v)(S_t) \quad \text{if nature uses } q \in \mathcal{Q}_S. \tag{19}$$

To prove equation 18, first condition on $\mathcal{G}_t$: by equation 17 and $p_t \in \mathcal{U}_{S_t A_t}$, the conditional expectation equals $r_{S_t A_t} + p_t^\top v \geq r_{S_t A_t} + m_{S_t A_t}(v)$. Conditioning on $\mathcal{F}_t$ and using $\mathbb{P}(A_t =$

$a \mid \mathcal{F}_t) = \pi(a \mid S_t)$ then gives $\sum_a \pi(a \mid S_t)\{r_{S_t a} + m_{S_t a}(v)\} = (T^\pi v)(S_t)$. The identity in equation 19 is equation 17 with $p_t = q_{S_t A_t}$, and the inequality holds because $A_t \in A(S_t)$. Both displays hold for arbitrary randomized history-dependent strategies of the opponent.

Lemma C.1. *The following statements hold for every $\pi \in \Pi_S$ and $0 < \epsilon < 1$.*

(i) The operators T, T^π, and H_q ($q \in \mathcal{Q}_S$) are order preserving, additively homogeneous, and nonexpansive in $\|\cdot\|_\infty$. Explicitly, each such operator F satisfies

$$x \le y \implies Fx \le Fy, \qquad F(x + c\mathbf{1}) = F(x) + c\mathbf{1}, \qquad \|Fx - Fy\|_\infty \le \|x - y\|_\infty$$

for all $x, y \in \mathbb{R}^S$ and $c \in \mathbb{R}$.

(ii) For $F \in \{T, T^\pi\}$, the map $x \mapsto F((1-\epsilon)x)$ is a $(1-\epsilon)$-contraction. Its unique fixed point is V_ϵ or V_ϵ^π, respectively, and has norm at most R/ϵ. Moreover, $V_\epsilon^\pi \le V_\epsilon$.

(iii) The vectors V_ϵ and V_ϵ^π are the discounted values against history-dependent opponents. In the control problem, both players have deterministic stationary discounted-optimal strategies that are optimal from every initial state; in particular, $V_\epsilon^{\pi_\epsilon} = V_\epsilon$ for some $\pi_\epsilon \in \Pi_D$. For fixed π, nature has such a strategy in $\mathcal{Q}_S$.

(iv) The N-horizon total values are $T^N 0$ and $(T^\pi)^N 0$. In particular, $N\, J_N(i; \pi, \tau) \ge [(T^\pi)^N 0]_i$ for every $i \in S$ and $\tau \in \mathcal{Q}_H$, and $N\, J_N(i; \sigma, q) \le [H_q^N 0]_i$ for every $q \in \mathcal{Q}_S$ and $\sigma \in \Pi_H$.

These are the rectangular dynamic-programming facts of Iyengar (2005) and Nilim & El Ghaoui (2004); see also Grand-Clement et al. (2023, Section 2.1 and Proposition 2.2). We include the verification because the row sets need not be convex and both players may randomize.

Proof. **(i) Operator properties.** Each coordinate of Tx, $T^\pi x$, or $H_q x$ is obtained from affine maps $x \mapsto r_{ia} + p^\top x$ with stochastic p by taking minima, maxima, and convex combinations. Each affine map is order preserving, commutes with the addition of $c\mathbf{1}$, and is 1-Lipschitz in $\|\cdot\|_\infty$. Minima, maxima, and convex combinations preserve these three properties.

(ii) Contraction and comparison. By (i), $\|F((1-\epsilon)x) - F((1-\epsilon)y)\|_\infty \le (1-\epsilon)\|x - y\|_\infty$, so Banach's fixed-point theorem gives a unique fixed point V. Since $\|F(0)\|_\infty \le R$,

$$\|V\|_\infty \le \|F((1-\epsilon)V) - F(0)\|_\infty + \|F(0)\|_\infty \le (1-\epsilon)\|V\|_\infty + R,$$

which gives $\|V\|_\infty \le R/\epsilon$. For the comparison, note that $T^\pi x \le Tx$ for every x, because a convex combination over actions is at most the maximum over actions. Hence $V_\epsilon = T((1-\epsilon)V_\epsilon) \ge T^\pi((1-\epsilon)V_\epsilon)$. Iterating the order-preserving contraction $x \mapsto T^\pi((1-\epsilon)x)$ from V_ϵ therefore produces a nonincreasing sequence, and its limit is the fixed point V_ϵ^π. Thus $V_\epsilon^\pi \le V_\epsilon$.

(iii) Discounted values. Write $V := V_\epsilon$. Choose $\pi_\epsilon(i) \in \arg\max_a \{r_{ia} + (1-\epsilon)m_{ia}(V)\}$ and, for every state-action pair, $q_{\epsilon,ia} \in \arg\min_{p \in \mathcal{U}_{ia}} p^\top V$. These choices exist because $A(i)$ is finite and $\mathcal{U}_{ia}$ is compact. Let nature use any $\tau \in \mathcal{Q}_H$ against π_ϵ. Since $T^{\pi_\epsilon}((1-\epsilon)V) = T((1-\epsilon)V) = V$, inequality equation 18 with $v = (1-\epsilon)V$ gives

$$\mathbb{E}\big[r_{S_t A_t} + (1-\epsilon)V(S_{t+1}) \;\big|\; \mathcal{F}_t\big] \ge V(S_t).$$

Multiply by $(1-\epsilon)^t$, take expectations, and sum over $t < N$. The V-terms telescope, and

$$\mathbb{E}_i \sum_{t=0}^{N-1} (1-\epsilon)^t r_{S_t A_t} \ge V_i - (1-\epsilon)^N \mathbb{E}_i V(S_N) \ge V_i - (1-\epsilon)^N R/\epsilon.$$

Letting $N \to \infty$ shows that the discounted payoff of (π_ϵ, τ) from i is at least V_i. Now let the controller use any $\sigma \in \Pi_H$ against q_ϵ. By equation 17,

$$\mathbb{E}\big[r_{S_t A_t} + (1-\epsilon)V(S_{t+1}) \;\big|\; \mathcal{G}_t\big] = r_{S_t A_t} + (1-\epsilon)m_{S_t A_t}(V) \le V(S_t),$$

and the same summation bounds the discounted payoff of (σ, q_ϵ) above by V_i. Hence V_ϵ is the value against history-dependent opponents, and $(\pi_\epsilon, q_\epsilon)$ is optimal from every initial state. Since $T^{\pi_\epsilon}((1-\epsilon)V_\epsilon) = V_\epsilon$ and the fixed point in (ii) is unique, also $V_\epsilon^{\pi_\epsilon} = V_\epsilon$. For fixed π, choose

$q_{ia}^{\pi,\epsilon} \in \arg\min_{p \in \mathcal{U}_{ia}} p^\top V_\epsilon^\pi$ for every (i,a). Inequality equation 18 with $v = (1-\epsilon)V_\epsilon^\pi$ gives the lower bound V_ϵ^π against every $\tau \in \mathcal{Q}_H$, and under $q^{\pi,\epsilon}$ the conditional expectation equals $V_\epsilon^\pi(S_t)$ exactly. Thus V_ϵ^π is nature's value, attained by $q^{\pi,\epsilon} \in \mathcal{Q}_S$ from every state.

(iv) Finite horizons. Put $W_k := T^k 0$. At time $t < N$, let the controller take an action attaining the maximum in $(TW_{N-t-1})(S_t)$. The proof of equation 18 uses only the conditional law of A_t given $\mathcal{F}_t$, so it applies to this time-dependent rule with $v = W_{N-t-1}$ and gives

$$\mathbb{E}\big[r_{S_t A_t} + W_{N-t-1}(S_{t+1}) \;\big|\; \mathcal{F}_t\big] \geq W_{N-t}(S_t)$$

against every nature strategy. Taking expectations and summing over $t < N$, the W-terms telescope and the N-stage total reward is at least $(W_N)_i$. If instead nature chooses rows minimizing $p^\top W_{N-t-1}$, the same computation gives the upper bound $(W_N)_i$ against every controller. Hence $T^N 0$ is the N-horizon value. Repeating the lower-bound computation with T^π gives $\mathbb{E}_i^{\pi,\tau} \sum_{t<N} r_{S_t A_t} \geq [(T^\pi)^N 0]_i$ for every $\tau \in \mathcal{Q}_H$, with equality when nature uses stagewise minimizing rows. Repeating the upper-bound computation for the singleton row sets $\{q_{ia}\}$, whose Bellman operator is H_q, gives $\mathbb{E}_i^{\sigma,q} \sum_{t<N} r_{S_t A_t} \leq [H_q^N 0]_i$ for every $\sigma \in \Pi_H$. □

Effective rows of a stationary controller. For $\pi \in \Pi_S$, put $r_i^\pi := \sum_a \pi(a \mid i) r_{ia}$ and

$$\mathcal{U}_i^\pi := \Big\{ \sum_{a \in A(i)} \pi(a \mid i) p_a : \; p_a \in \mathcal{U}_{ia} \text{ for all } a \Big\}, \qquad (T^\pi x)_i = r_i^\pi + \min_{\bar{p} \in \mathcal{U}_i^\pi} \bar{p}^\top x. \tag{20}$$

The set $\mathcal{U}_i^\pi$ is compact because it is the image of the compact product $\prod_a \mathcal{U}_{ia}$ under a continuous map. The identity for T^π holds because the minimum of $\sum_a \pi(a \mid i) p_a^\top x$ over tuples $(p_a)_a$ is attained by minimizing each positive-weight component separately. A full selector $q \in \mathcal{Q}_S$ induces the effective rows $\sum_a \pi(a \mid i) q_{ia} \in \mathcal{U}_i^\pi$. Conversely, every choice of effective rows is induced by some $q \in \mathcal{Q}_S$: choose a representing tuple at each state and assign arbitrary feasible rows to actions of probability zero. Stationary quantities of the fixed-policy problem, such as $\eta^{\pi,q}$ and canonical biases, therefore range over the same set of transition matrices whether nature is described by full selectors or by effective rows.

C.2 Finite-chain facts used by the Bellman arguments

For a finite stochastic matrix P and reward vector c, define

$$P^\infty := \lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} P^t, \quad Z_P := (I - P + P^\infty)^{-1}, \quad \eta := P^\infty c, \quad w := Z_P(c - \eta). \tag{21}$$

Lemma C.2. *The projector P^∞ is stochastic, $PP^\infty = P^\infty P = (P^\infty)^2 = P^\infty$, and Z_P exists. The canonical bias w is the unique solution of*

$$(I - P)w = c - P^\infty c, \qquad P^\infty w = 0. \tag{22}$$

If $d \geq 0$ and $P^\infty d = 0$, then d vanishes on recurrent states and

$$Z_P d = \sum_{t=0}^{\infty} P^t d \geq 0. \tag{23}$$

Proof. **Projector identities.** The existence of the Cesàro limit, its stochasticity, the identities $PP^\infty = P^\infty P = (P^\infty)^2 = P^\infty$, and the invertibility of $I - P + P^\infty$ are standard for finite chains and hold without irreducibility or aperiodicity (Schweitzer, 1985, Section 2, equations (2.2)–(2.9)), (Puterman, 2014, Chapter 8).

Poisson system. Multiplying $Z_P^{-1} = I - P + P^\infty$ by P^∞ on the left gives $P^\infty Z_P^{-1} = P^\infty$, hence $P^\infty Z_P = P^\infty$. Since P commutes with Z_P^{-1}, it commutes with Z_P, and

$$(I - P)Z_P = (Z_P^{-1} - P^\infty)Z_P = I - P^\infty.$$

For $w := Z_P(c - P^\infty c)$ these identities give $(I - P)w = (I - P^\infty)(c - P^\infty c) = c - P^\infty c$ and $P^\infty w = P^\infty(c - P^\infty c) = 0$. For uniqueness, suppose $(I - P)x = 0$ and $P^\infty x = 0$. Then $x = P^t x$ for every t, so averaging over t gives $x = P^\infty x = 0$.

Transient series. The invariant distribution of each recurrent class is strictly positive on that class. The coordinate of $P^\infty d$ at a recurrent state is the invariant mean of d over its class. This mean is zero and $d \geq 0$, so d vanishes on every recurrent class. Let Q be the transient block of P and d_{tr} the restriction of d to the transient states. Recurrent classes are closed, so $P^t d$ vanishes on recurrent states and equals $Q^t d_{\mathrm{tr}}$ on transient states. Since $\rho(Q) < 1$, the vector $v := \sum_{t \geq 0} P^t d$ is finite and equals $(I - Q)^{-1} d_{\mathrm{tr}} \geq 0$ on the transient states and zero elsewhere. It satisfies $(I - P)v = d$. It also satisfies $P^\infty v = 0$, because v vanishes on recurrent states and P^∞ has zero columns at transient states. Uniqueness in equation 22, applied with $c = d$ and $P^\infty d = 0$, gives $v = Z_P d$. □

Lemma C.3. *For each fixed finite chain, $\epsilon(I - (1 - \epsilon)P)^{-1} c \to P^\infty c$ as $\epsilon \downarrow 0$. Under a stationary pair (π, q), X_N converges almost surely to the invariant mean reward of the recurrent class eventually entered. From initial state i, the expectation of this limiting random variable is $\eta_i^{\pi,q}$. All four payoffs in equation 1 equal this number.*

Proof. For a finite stochastic matrix, the Cesàro limit P^∞ is also the Abel limit, which is the first claim. Under (π, q) the state process is a finite Markov chain with matrix $P^{\pi,q}$. It enters a recurrent class almost surely, and the ergodic theorem for that class identifies the almost-sure limit of $N^{-1} \sum_{t<N} r^\pi_{S_t}$. The differences $r_{S_t A_t} - r^\pi_{S_t}$ are martingale differences with respect to $(\mathcal{F}_{t+1})$ and are bounded by $2R$, so their averages tend to zero almost surely by the strong law for martingales with bounded increments. Hence X_N has the stated limit. Since $|X_N| \leq R$, bounded convergence gives $\mathbb{E}_i X_N \to \eta_i^{\pi,q}$ and identifies all four payoffs. These are finite-chain statements, and periodic recurrent classes are allowed (Puterman, 2014, Chapters 8–9). □

C.3 Normalized finite-horizon values

Lemma C.4. *Let $F \in \{T, T^\pi, H_q\}$ with $\pi \in \Pi_S$ and $q \in \mathcal{Q}_S$, and let V_ϵ^F denote the fixed point of $x \mapsto F((1 - \epsilon)x)$. If ϵV_ϵ^F converges in $\|\cdot\|_\infty$ as $\epsilon \downarrow 0$, then $F^N 0/N$ converges to the same limit.*

Proof. We verify the hypotheses of Ziliotto (2016, Theorem 1.2) on the Banach space $(\mathbb{R}^S, \|\cdot\|_\infty)$. The map F is nonexpansive by Lemma C.1(i). For $\lambda, \mu \in (0, 1]$ and $x \in \mathbb{R}^S$,

$$\|\lambda F(x/\lambda) - \mu F(x/\mu)\|_\infty \leq R|\lambda - \mu|. \tag{24}$$

Indeed, $\lambda(T(x/\lambda))_i = \max_a \{\lambda r_{ia} + m_{ia}(x)\}$, and changing λ to μ changes each expression inside the maximum by at most $R|\lambda - \mu|$. The same computation applies to H_q, and for T^π the difference is exactly $(\lambda - \mu) r^\pi$. Hence Assumption 1 of that reference holds with constant R; if $R = 0$, any positive constant works. In that reference, the discounted family is the solution of $v_\lambda = \lambda F((1 - \lambda)v_\lambda/\lambda)$, which is $v_\epsilon = \epsilon V_\epsilon^F$ at $\lambda = \epsilon$, and the finite-horizon sequence is $v_N = F^N 0/N$. Theorem 1.2 of that reference states that one family converges if and only if the other does, with the same limit. No definability assumption enters equation 24. □

C.4 Payoff ordering and a pathwise comparison principle

For bounded rewards, Fatou's lemma and its reverse form give

$$I_i^- \leq J_i^- \leq J_i^+ \leq I_i^+. \tag{25}$$

The next lemma supplies the two pathwise endpoints of this chain. It uses only normalized discounted values and a one-sided recession inequality. It does not require a finite bias or stationary attainment.

Lemma C.5 (Pathwise comparison)**.** *Fix $u \in \mathbb{R}^S$ and an initial state i.*

(a) Let $\pi \in \Pi_S$ and suppose that

$$u_j \leq \sum_a \pi(a \mid j)\, m_{ja}(u) \quad (j \in S), \qquad \liminf_{\epsilon \downarrow 0} \min_{j \in S} \big(\epsilon V_\epsilon^\pi - u\big)_j \geq 0. \tag{26}$$

Then, for every $\tau \in \mathcal{Q}_H$, under (π, τ) the sequence $u(S_t)$ converges almost surely to a random variable U_∞ with $\mathbb{E}_i U_\infty \geq u_i$, and $\liminf_N X_N \geq U_\infty$ almost surely. In particular, $I_i^-(\pi, \tau) \geq u_i$.

(b) Let $q \in \mathcal{Q}_S$, and let V_ϵ^q be the fixed point of $x \mapsto H_q((1-\epsilon)x)$. Suppose that

$$q_{ja}^\top u \le u_j \quad (j \in S,\ a \in A(j)), \qquad \limsup_{\epsilon \downarrow 0} \max_{j \in S} \big(\epsilon V_\epsilon^q - u\big)_j \le 0. \tag{27}$$

Then, for every $\sigma \in \Pi_H$, under (σ, q) the sequence $u(S_t)$ converges almost surely to a random variable U_∞ with $\mathbb{E}_i U_\infty \le u_i$, and $\limsup_N X_N \le U_\infty$ almost surely. In particular, $I_i^+(\sigma, q) \le u_i$.

Proof. We first prove (a), keeping the nature strategy τ fixed throughout.

Step 1: compare rewards with a discounted value. Fix $\epsilon \in (0,1)$ and set $V := V_\epsilon^\pi$ and $w := \epsilon V$. Define the one-step discrepancy

$$\zeta_t := r_{S_t A_t} + (1-\epsilon)V(S_{t+1}) - V(S_t), \qquad b_t := \mathbb{E}[\zeta_t \mid \mathcal{F}_t].$$

The fixed-point equation and equation 18 give $b_t \ge 0$. Solving the definition of ζ_t for the reward and summing yields the exact identity

$$\sum_{t=0}^{N-1} r_{S_t A_t} = V(S_0) - (1-\epsilon)V(S_N) + \sum_{t=1}^{N-1} w(S_t) + \sum_{t=0}^{N-1} \zeta_t.$$

The first two terms form a bounded boundary contribution. The middle sum is the quantity that will approach the long-run gain.

Step 2: show that random fluctuations have zero time average. Set $D_t := \zeta_t - b_t$. Then D_t is measurable with respect to $\mathcal{F}_{t+1}$ and satisfies $\mathbb{E}[D_t \mid \mathcal{F}_t] = 0$. Thus its partial sums form a martingale. Since $\|V\|_\infty \le R/\epsilon$, the increments obey the deterministic bound $|D_t| \le 2(R + 2R/\epsilon)$. For completeness, if $|D_t| \le C_\epsilon$ and $C_\epsilon > 0$, the Azuma–Hoeffding inequality gives, for each $a > 0$,

$$\mathbb{P}\left(\left|\sum_{t<N} D_t\right| \ge aN\right) \le 2\exp\left(-\frac{a^2 N}{2C_\epsilon^2}\right).$$

The sum of these probabilities over N is finite. Borel–Cantelli, applied for positive rational a, therefore gives $N^{-1}\sum_{t<N} D_t \to 0$ almost surely. If $C_\epsilon = 0$, the conclusion is immediate. Since $\zeta_t = D_t + b_t \ge D_t$, division of the identity in Step 1 by N gives

$$\liminf_{N\to\infty} X_N \ge \liminf_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} w(S_t) \qquad \text{almost surely.}$$

Here the boundary contribution and the single added term $w(S_0)/N$ vanish because ϵ is fixed while $N \to \infty$.

Step 3: replace the discounted value by u. The second hypothesis in equation 26 allows us to choose a decreasing sequence $\epsilon_k \downarrow 0$ such that $\epsilon_k V_{\epsilon_k}^\pi \ge u - k^{-1}\mathbf{1}$. For each fixed k, Step 2 gives

$$\liminf_{N} X_N \ge \liminf_{N} \frac{1}{N}\sum_{t<N} u(S_t) - \frac{1}{k} \qquad \text{almost surely.}$$

Take the intersection of these countably many probability-one events, and then let $k \to \infty$. This proves

$$\liminf_{N} X_N \ge \liminf_{N} \frac{1}{N}\sum_{t<N} u(S_t) \qquad \text{almost surely.}$$

This argument takes the time limit first for each fixed discount. It does not require a bound on V_ϵ that is uniform as $\epsilon \downarrow 0$.

Step 4: identify the limiting continuation gain. By the transition rule and the first hypothesis in equation 26,

$$\mathbb{E}[u(S_{t+1}) \mid \mathcal{F}_t] \ge \sum_a \pi(a \mid S_t) m_{S_t a}(u) \ge u(S_t).$$

Thus $u(S_t)$ is a submartingale: its conditional expected next value is at least its present value. It is bounded because S is finite. The bounded submartingale convergence theorem and bounded convergence give a random variable U_∞ such that

$$u(S_t) \longrightarrow U_\infty \quad \text{almost surely and in } L^1, \qquad \mathbb{E}_i U_\infty = \lim_t \mathbb{E}_i u(S_t) \geq u_i.$$

The time averages of a convergent sequence have the same limit. Step 3 consequently gives $\liminf_N X_N \geq U_\infty$ almost surely. Taking expectations proves $I_i^-(\pi, \tau) \geq u_i$.

Part (b). Fix $\sigma \in \Pi_H$ and repeat Steps 1–2 with $V := V_\epsilon^q$ and $w := \epsilon V$. The nominal discounted fixed point exists and satisfies $\|V\|_\infty \leq R/\epsilon$ by Lemma C.1, applied to singleton row sets. Inequality equation 19 first gives $\mathbb{E}[\zeta_t \mid \mathcal{G}_t] \leq 0$. The tower property then gives $b_t := \mathbb{E}[\zeta_t \mid \mathcal{F}_t] \leq 0$. With $D_t := \zeta_t - b_t$, we now have $\zeta_t \leq D_t$. Therefore

$$\limsup_N X_N \leq \limsup_N \frac{1}{N} \sum_{t<N} w(S_t) \qquad \text{almost surely.}$$

Choose $\epsilon_k \downarrow 0$ such that $\epsilon_k V_{\epsilon_k}^q \leq u + k^{-1}\mathbf{1}$, and use the countable-intersection argument from Step 3 to replace w by u. Finally,

$$\mathbb{E}[u(S_{t+1}) \mid \mathcal{F}_t] = \mathbb{E}[q_{S_t A_t}^\top u \mid \mathcal{F}_t] \leq u(S_t).$$

Hence $u(S_t)$ is a bounded supermartingale. It converges almost surely and in L^1 to U_∞ with $\mathbb{E}_i U_\infty \leq u_i$. Its time average has the same limit, so $\limsup_N X_N \leq U_\infty$ almost surely and $I_i^+(\sigma, q) \leq u_i$. □

C.5 The nominal MDP obtained by fixing nature

Lemma C.6. *Fix $q \in \mathcal{Q}_S$ and define $d_i(q) := \max_{\pi \in \Pi_D} \eta_i^{\pi,q}$, the optimal gain of the nominal MDP obtained by fixing q.*

(i) $\epsilon V_\epsilon^q \to d(q)$*, and one policy $\pi_q \in \Pi_D$ satisfies $\eta^{\pi_q, q} = d(q)$ at every state.*

(ii) $d_i(q) = \max_{a \in A(i)} q_{ia}^\top d(q)$ *for every i.*

(iii) For every payoff $\Psi \in \{I^-, J^-, J^+, I^+\}$ and every state, $\sup_{\sigma \in \Pi_H} \Psi_i(\sigma, q) = d_i(q)$. In particular, $\max_{\pi \in \Pi_D} \mu^\top \eta^{\pi,q} = \mu^\top d(q)$ for every $\mu \in \Delta(S)$.

(iv) $H_q^N 0/N \to d(q)$*. Consequently, for every $\delta > 0$ there is $N_\delta(q) < \infty$ such that $J_N(i; \sigma, q) \leq d_i(q) + \delta$ for all $i \in S$, $\sigma \in \Pi_H$, and $N \geq N_\delta(q)$.*

Proof. **(i) Discounted limit and a common optimizer.** For $\pi \in \Pi_D$, the fixed-q discounted value is $V_\epsilon^{\pi,q} = (I - (1-\epsilon)P^{\pi,q})^{-1} r^\pi$, and $\epsilon V_\epsilon^{\pi,q} \to \eta^{\pi,q}$ by Lemma C.3. Lemma C.1, applied to the singleton row sets $\{q_{ia}\}$, gives $V_\epsilon^q \geq V_\epsilon^{\pi,q}$ for every $\pi \in \Pi_D$, with equality for a greedy $\pi_\epsilon \in \Pi_D$. Hence $\epsilon V_\epsilon^q = \max_{\pi \in \Pi_D} \epsilon V_\epsilon^{\pi,q}$, a finite maximum of convergent families, and $\epsilon V_\epsilon^q \to d(q)$. Take $\epsilon_k \downarrow 0$. Because Π_D is finite, some π_q equals π_{ϵ_k} for infinitely many k. Along those indices, $\epsilon_k V_{\epsilon_k}^{\pi_q, q} = \epsilon_k V_{\epsilon_k}^q$, and the two sides converge to $\eta^{\pi_q, q}$ and $d(q)$, respectively.

(ii) Recession equation. Put $v_\epsilon := \epsilon V_\epsilon^q$. The fixed-point equation gives $(v_\epsilon)_i = \max_a \{\epsilon r_{ia} + (1-\epsilon) q_{ia}^\top v_\epsilon\}$. Comparing each term with $q_{ia}^\top d(q)$ gives

$$\left| (v_\epsilon)_i - \max_a q_{ia}^\top d(q) \right| \leq \epsilon R + (1-\epsilon)\|v_\epsilon - d(q)\|_\infty + \epsilon \|d(q)\|_\infty.$$

The right side tends to zero, and $v_\epsilon \to d(q)$ by (i).

(iii) All four payoffs. By (i) and (ii), $u := d(q)$ satisfies equation 27. Lemma C.5(b) gives $I_i^+(\sigma, q) \leq d_i(q)$ for every $\sigma \in \Pi_H$, and equation 25 extends this upper bound to every Ψ. Conversely, $\Psi_i(\pi_q, q) = \eta_i^{\pi_q, q} = d_i(q)$ by Lemma C.3. For the weighted identity, $\eta^{\pi,q} \leq d(q)$ for every $\pi \in \Pi_D$ gives $\max_\pi \mu^\top \eta^{\pi,q} \leq \mu^\top d(q)$, and the policy π_q attains equality.

(iv) Finite horizons. Lemma C.4 with $F = H_q$ and part (i) give $H_q^N 0/N \to d(q)$. Finite-horizon dynamic programming for the nominal MDP (Lemma C.1(iv) with singleton row sets) gives $N\, J_N(i; \sigma, q) \leq [H_q^N 0]_i$ for every $\sigma \in \Pi_H$. Choose $N_\delta(q)$ such that $\|H_q^N 0/N - d(q)\|_\infty \leq \delta$ for $N \geq N_\delta(q)$. □

C.6 WEIGHTED STATIONARY DUALITY

The following lemma is the stationary duality of Grand-Clement et al. (2023, Theorem 3.5, equation (3.6)), stated for an arbitrary initial law. We give a proof whose finite restriction keeps the selected rows themselves, so that convex row sets are not needed.

Lemma C.7. *For every initial distribution $\mu \in \Delta(S)$,*

$$\inf_{q\in\mathcal{Q}_S}\max_{\pi\in\Pi_D}\mu^\top\eta^{\pi,q}=\max_{\pi\in\Pi_D}\inf_{q\in\mathcal{Q}_S}\mu^\top\eta^{\pi,q}. \tag{28}$$

Proof. Weak duality gives the inequality "$\geq$": for every controller and selector, the maximum over controllers is at least their pairwise payoff, which is at least the infimum over selectors. We prove the opposite inequality by retaining finitely many feasible rows.

Step 1: retain one nearly worst reply to each deterministic controller. Fix $\xi > 0$. Since Π_D is finite and rewards are bounded, for each $\pi \in \Pi_D$ we can choose $q^\pi \in \mathcal{Q}_S$ such that

$$\mu^\top\eta^{\pi,q^\pi}\leq\inf_{q\in\mathcal{Q}_S}\mu^\top\eta^{\pi,q}+\xi.$$

Define $E_{ia} := \{q^\pi_{ia} : \pi \in \Pi_D\}$ and $\mathcal{Q}_E := \prod_{i,a} E_{ia}$. These sets are finite, $\mathcal{Q}_E \subseteq \mathcal{Q}_S$, and every selected reply q^π belongs to $\mathcal{Q}_E$.

Step 2: obtain a stationary saddle in the finite restriction. Restrict nature to rows in E_{ia}. For each discount parameter, Lemma C.1(iii) supplies a discounted saddle in $\Pi_D \times \mathcal{Q}_E$. Both sets are finite. Along a sequence $\epsilon_k \downarrow 0$, we can therefore keep the entire discounted-optimal pair constant, say $(\bar\pi, \bar q)$. Its discounted saddle inequalities, restricted to stationary opponents, are

$$V^{\pi,\bar q}_{\epsilon_k}\leq V^{\bar\pi,\bar q}_{\epsilon_k}\leq V^{\bar\pi,q}_{\epsilon_k}\qquad(\pi\in\Pi_D,\ q\in\mathcal{Q}_E).$$

Multiplication by ϵ_k and Lemma C.3 give

$$\eta^{\pi,\bar q}\leq v^E\leq\eta^{\bar\pi,q},\qquad v^E:=\eta^{\bar\pi,\bar q}.$$

These comparisons hold at every initial state. After averaging with μ, they imply

$$\min_{q\in\mathcal{Q}_E}\max_{\pi\in\Pi_D}\mu^\top\eta^{\pi,q}\leq\mu^\top v^E\leq\max_{\pi\in\Pi_D}\min_{q\in\mathcal{Q}_E}\mu^\top\eta^{\pi,q}.$$

Only a saddle against stationary opponents is needed here, so the finite-chain limit is sufficient.

Step 3: return to the original row sets. Using $\mathcal{Q}_E \subseteq \mathcal{Q}_S$ for the first inequality, Step 2 for the second, and $q^\pi \in \mathcal{Q}_E$ for the third, we obtain

$$\begin{aligned}\inf_{q\in\mathcal{Q}_S}\max_{\pi\in\Pi_D}\mu^\top\eta^{\pi,q}&\leq\min_{q\in\mathcal{Q}_E}\max_{\pi\in\Pi_D}\mu^\top\eta^{\pi,q}\\&\leq\max_{\pi\in\Pi_D}\min_{q\in\mathcal{Q}_E}\mu^\top\eta^{\pi,q}\\&\leq\max_{\pi\in\Pi_D}\mu^\top\eta^{\pi,q^\pi}\\&\leq\max_{\pi\in\Pi_D}\inf_{q\in\mathcal{Q}_S}\mu^\top\eta^{\pi,q}+\xi.\end{aligned}$$

Letting $\xi \downarrow 0$ proves the required inequality. The construction uses actual rows from each $\mathcal{U}_{ia}$, so it also applies when the row sets are nonconvex. □

D PROOF OF THEOREM 3.1: POLICY EVALUATION

Throughout this section $\pi \in \Pi_S$ is fixed. The proof has three steps. Lemma D.1 identifies the normalized discounted limit and constructs one nearly worst stationary kernel for all initial states. Proposition D.2 turns the discounted limit into pathwise payoff bounds. Proposition D.3 adds finite-horizon convergence and the uniform finite-horizon guarantee.

Lemma D.1. *For every $\pi \in \Pi_S$ the following hold.*

(i) The normalized discounted values converge:

$$\epsilon V^\pi_\epsilon\longrightarrow g^\pi,\qquad g^\pi_i:=\inf_{q\in\mathcal{Q}_S}\eta^{\pi,q}_i. \tag{29}$$

(ii) For every $\nu > 0$, one full selector $q^{\pi,\nu} \in \mathcal{Q}_S$ satisfies

$$g^\pi \le \eta^{\pi,q^{\pi,\nu}} \le g^\pi + \nu\mathbf{1}. \tag{30}$$

(iii) The gain satisfies the fixed-policy recession equation

$$g_i^\pi = \sum_a \pi(a \mid i)\, m_{ia}(g^\pi) \qquad (i \in S). \tag{31}$$

Proof. **(i) Discounted limit.** The hypotheses of Grand-Clement et al. (2023, Lemma 4.7) are satisfied: the state and controller action sets are finite, the ambiguity set $\prod_{i,a} \mathcal{U}_{ia}$ is nonempty, compact, and (s,a)-rectangular, and π is stationary. That lemma allows randomized stationary π and does not require convex row sets. Its reward model includes ours by taking the reward independent of the successor state, and its discount factor is $1-\epsilon$. With initial law e_i, the lemma states that $\min_{q\in\mathcal{Q}_S} \epsilon(V_\epsilon^{\pi,q})_i$ converges to $\inf_{q\in\mathcal{Q}_S} \eta_i^{\pi,q}$ as $\epsilon \downarrow 0$. By Lemma C.1(iii), this minimum equals $(\epsilon V_\epsilon^\pi)_i$. Hence equation 29 holds coordinatewise, and in supremum norm because S is finite.

(ii) One selector for all states. The same ambiguity set is also s-rectangular: group the rows at state i into the compact set $\prod_a \mathcal{U}_{ia}$. Its stationary kernels are exactly the full selectors in $\mathcal{Q}_S$, so Grand-Clement et al. (2023, Theorem 4.3) applies to the fixed stationary controller π. Let $\mu_i := 1/n$ be the uniform initial law. Grand-Clement et al. (2023, Theorem 4.3), applied with this law and tolerance ν/n, gives $q^{\pi,\nu} \in \mathcal{Q}_S$ and $\epsilon_\nu > 0$ such that, for all $\epsilon \in (0, \epsilon_\nu)$,

$$0 \le \epsilon\mu^\top \big(V_\epsilon^{\pi,q^{\pi,\nu}} - V_\epsilon^\pi\big) \le \nu/n.$$

Here we used that the reference quantity $\min_{q\in\mathcal{Q}_S} \epsilon\mu^\top V_\epsilon^{\pi,q}$ of that theorem equals $\epsilon\mu^\top V_\epsilon^\pi$, because one selector attains V_ϵ^π at every state (Lemma C.1(iii)); the same fact gives the left inequality. Every coordinate of the nonnegative vector $\epsilon(V_\epsilon^{\pi,q^{\pi,\nu}} - V_\epsilon^\pi)$ is at most n times its μ-average, so

$$0 \le \epsilon\big(V_\epsilon^{\pi,q^{\pi,\nu}} - V_\epsilon^\pi\big) \le \nu\mathbf{1} \qquad (0 < \epsilon < \epsilon_\nu). \tag{32}$$

Keep the selector fixed and let $\epsilon \downarrow 0$. Lemma C.3 gives $\epsilon V_\epsilon^{\pi,q^{\pi,\nu}} \to \eta^{\pi,q^{\pi,\nu}}$, and (i) gives $\epsilon V_\epsilon^\pi \to g^\pi$. Hence $\eta^{\pi,q^{\pi,\nu}} \le g^\pi + \nu\mathbf{1}$. The lower bound in equation 30 is the definition of g^π as an infimum.

(iii) Recession equation. Put $u_\epsilon := \epsilon V_\epsilon^\pi$. The fixed-point equation $V_\epsilon^\pi = T^\pi((1-\epsilon)V_\epsilon^\pi)$ reads $(u_\epsilon)_i = \epsilon r_i^\pi + (1-\epsilon)\sum_a \pi(a \mid i) m_{ia}(u_\epsilon)$. Since m_{ia} is 1-Lipschitz,

$$\Big|(u_\epsilon)_i - \sum_a \pi(a \mid i) m_{ia}(g^\pi)\Big| \le \epsilon R + (1-\epsilon)\|u_\epsilon - g^\pi\|_\infty + \epsilon\|g^\pi\|_\infty.$$

The right side tends to zero by (i), and $u_\epsilon \to g^\pi$. This proves equation 31. □

Remark 7 (Convex hulls)**.** *The cited results already allow nonconvex row sets. The following separate argument explains why convexification also preserves the fixed-policy average value and the existence of common approximate replies. Convexification changes neither the discounted operators nor V_ϵ^π, so the limit g^π in (i) is unchanged and equals the infimum of $\eta^{\pi,\tilde{q}}$ over convexified selectors $\tilde{q}$. Since $\mathcal{Q}_S \subseteq \prod_{i,a} \mathrm{co}(\mathcal{U}_{ia})$, the infimum over original selectors is at least g^π. The reverse inequality and part (ii) follow from a replacement argument. Let $\tilde{q}$ be a convexified selector with $\eta^{\pi,\tilde{q}} \le g^\pi + \nu\mathbf{1}$, obtained by running the proof of (ii) in the convexified model. By Carathéodory's theorem, each row $\tilde{q}_{ia}$ is a convex combination $\sum_k \lambda_{iak} p_{iak}$ of finitely many rows $p_{iak} \in \mathcal{U}_{ia}$. With only these rows available, nature faces a finite MDP whose action at state i is a tuple $(p_a)_a$ of listed rows, with transition row $\sum_a \pi(a \mid i) p_a$. Drawing the components independently with laws $(\lambda_{iak})_k$ is a randomized stationary policy of this finite MDP, and by linearity it induces the chain $P^{\pi,\tilde{q}}$. A deterministic stationary policy of this finite MDP minimizes the gain at every state simultaneously (Puterman, 2014, Chapter 9). It is a selector $q \in \mathcal{Q}_S$ with $\eta^{\pi,q} \le \eta^{\pi,\tilde{q}} \le g^\pi + \nu\mathbf{1}$.*

Proposition D.2. *For every $i \in S$,*

$$\begin{aligned} \inf_{\tau\in\mathcal{Q}_H} I_i^-(\pi,\tau) &= \inf_{q\in\mathcal{Q}_S} \eta_i^{\pi,q} = g_i^\pi \qquad (\pi \in \Pi_S), \\ \sup_{\sigma\in\Pi_H} I_i^+(\sigma,q) &= \max_{\pi\in\Pi_D} \eta_i^{\pi,q} = d_i(q) \quad (q \in \mathcal{Q}_S). \end{aligned} \tag{33}$$

Proof. **First line.** By Lemma D.1(i) and (iii), the vector $u := g^\pi$ satisfies both conditions in equation 26. Lemma C.5(a) therefore gives $I_i^-(\pi,\tau) \geq g_i^\pi$ for every $\tau \in \mathcal{Q}_H$. Conversely, $\mathcal{Q}_S \subseteq \mathcal{Q}_H$ and $I_i^-(\pi,q) = \eta_i^{\pi,q}$ for stationary q (Lemma C.3), so the infimum over $\mathcal{Q}_H$ is at most $\inf_q \eta_i^{\pi,q} = g_i^\pi$.

Second line. This is Lemma C.6(iii) with $\Psi = I^+$. □

These identities are pathwise counterparts of Grand-Clement et al. (2023, Lemma 3.3), which treats the expected limit superior. The proof above uses only normalized discounted values, so it covers the expected limit inferior and randomized controllers directly.

Proposition D.3. *For each $\pi \in \Pi_S$, $(T^\pi)^N 0/N \to g^\pi$. Consequently, for every $\delta > 0$ there is $N_\delta < \infty$ such that $J_N(i;\pi,\tau) \geq g_i^\pi - \delta$ for all $i \in S$, $N \geq N_\delta$, and $\tau \in \mathcal{Q}_H$. For every $\Psi \in \{I^-, J^-, J^+, I^+\}$ and every state, the stationary and history-dependent infima coincide:*

$$\inf_{\tau \in \mathcal{Q}_H} \Psi_i(\pi,\tau) = \inf_{q \in \mathcal{Q}_S} \Psi_i(\pi,q) = g_i^\pi.$$

Proof. **Step 1: finite-horizon limit.** Lemma D.1(i) and Lemma C.4 with $F = T^\pi$ give $(T^\pi)^N 0/N \to g^\pi$.

Step 2: uniform finite-horizon guarantee. By Lemma C.1(iv), $J_N(i;\pi,\tau) \geq [(T^\pi)^N 0]_i/N$ for every $\tau \in \mathcal{Q}_H$. Choose N_δ such that $\|(T^\pi)^N 0/N - g^\pi\|_\infty \leq \delta$ for $N \geq N_\delta$. The resulting bound holds for all states and opponents at once.

Step 3: the four payoffs. Fix Ψ, i, and $\nu > 0$. Then

$$g_i^\pi = \inf_\tau I_i^-(\pi,\tau) \leq \inf_\tau \Psi_i(\pi,\tau) \leq \inf_q \Psi_i(\pi,q) \leq \Psi_i(\pi,q^{\pi,\nu}) = \eta_i^{\pi,q^{\pi,\nu}} \leq g_i^\pi + \nu.$$

The equality on the left is Proposition D.2. The first inequality is equation 25. The second holds because $\mathcal{Q}_S \subseteq \mathcal{Q}_H$. The equality on the right is Lemma C.3, and the last inequality is equation 30. Letting $\nu \downarrow 0$ completes the proof of Theorem 3.1. □

E Proof of Theorem 3.2: uniform control

Theorem E.1. *The limits and the common controller in equation 4 exist, and they satisfy the uniform bounds equation 5.*

Proof. The proof has four steps. Steps 1–2 treat the controller and Steps 3–4 treat nature. Recall $d(q)$ and the operator H_q from Lemma C.6.

Step 1: the optimal gain and one optimal controller. Define $\bar{g}_i := \max_{\pi \in \Pi_D} g_i^\pi$. Lemma C.1(ii)–(iii) gives the coordinatewise identity

$$\epsilon V_\epsilon = \max_{\pi \in \Pi_D} \epsilon V_\epsilon^\pi.$$

Indeed, every fixed-policy value is at most V_ϵ, and one deterministic discounted-optimal controller attains equality at all states. Since Π_D is finite and every ϵV_ϵ^π converges to g^π,

$$\|\epsilon V_\epsilon - \bar{g}\|_\infty \leq \max_{\pi \in \Pi_D} \|\epsilon V_\epsilon^\pi - g^\pi\|_\infty \longrightarrow 0.$$

To obtain one average-optimal controller, take any sequence $\epsilon_k \downarrow 0$ and a deterministic discounted-optimal controller at each ϵ_k. Finiteness of Π_D supplies a subsequence on which this controller is constant, say $\pi^\star$. Along that subsequence,

$$\epsilon_k V_{\epsilon_k}^{\pi^\star} = \epsilon_k V_{\epsilon_k} \quad \Longrightarrow \quad g^{\pi^\star} = \bar{g}.$$

Set $g^\star := \bar{g}$. Thus the coordinatewise maximum is attained by the same controller at every state. This also gives the vector version of Grand-Clement et al. (2023, Lemma 4.8).

Step 2: finite horizons and the controller's guarantee. Lemma C.4 with $F = T$ and Step 1 give $T^N 0/N \to g^\star$, which completes equation 4. Fix $\delta > 0$. Proposition D.3 applied to $\pi^\star$ gives $N_1 < \infty$ such that $J_N(i;\pi^\star,\tau) \geq g_i^\star - \delta$ for all i, all $\tau \in \mathcal{Q}_H$, and all $N \geq N_1$. This is the lower inequality in equation 5.

Step 3: one nearly worst selector for nature. Let $\mu_i := 1/n$. We evaluate both sides of equation 28. By Lemma C.6(iii), $\max_{\pi\in\Pi_D} \mu^\top \eta^{\pi,q} = \mu^\top d(q)$ for every q. By Lemma D.1(ii), $\inf_q \mu^\top \eta^{\pi,q} = \mu^\top g^\pi$ for every π. By Step 1, $\max_{\pi\in\Pi_D} \mu^\top g^\pi = \mu^\top g^\star$. Lemma C.7 therefore gives

$$\mu^\top g^\star = \max_{\pi\in\Pi_D} \inf_{q\in\mathcal{Q}_S} \mu^\top \eta^{\pi,q} = \inf_{q\in\mathcal{Q}_S} \mu^\top d(q). \tag{34}$$

Moreover, $d(q) \geq \eta^{\pi^\star,q} \geq g^{\pi^\star} = g^\star$ for every $q \in \mathcal{Q}_S$. Given $\alpha > 0$, choose q_α with $\mu^\top(d(q_\alpha) - g^\star) \leq \alpha/n$. The vector $d(q_\alpha) - g^\star$ is nonnegative, and each of its coordinates is at most n times its μ-average. Hence

$$g^\star \leq d(q_\alpha) \leq g^\star + \alpha\mathbf{1}. \tag{35}$$

Step 4: nature's finite-horizon guarantee. Apply Lemma C.6(iv) to $q_{\delta/2}$. It gives $N_2 < \infty$ such that, for all i, all $\sigma \in \Pi_H$, and all $N \geq N_2$,

$$J_N(i;\sigma,q_{\delta/2}) \leq d_i(q_{\delta/2}) + \delta/2 \leq g_i^\star + \delta.$$

With the selector $q_{\delta/2}$ in the role of q_δ in equation 5 and $N_\delta := \max\{N_1, N_2\}$, both inequalities of equation 5 hold simultaneously. □

Two selection arguments turn statewise value identities into all-state guarantees. The constant-policy subsequence in Step 1 produces one optimal controller, and the full-support law in Step 3 produces one nearly worst kernel.

F Payoff conventions and strategy-class duality

Theorem F.1. *Let $\Pi_D \subseteq \mathcal{C} \subseteq \Pi_H$ and $\mathcal{Q}_S \subseteq \mathcal{N} \subseteq \mathcal{Q}_H$. For every $\Psi \in \{I^-, J^-, J^+, I^+\}$ and $i \in S$,*

$$\sup_{\sigma\in\mathcal{C}} \inf_{\tau\in\mathcal{N}} \Psi_i(\sigma,\tau) = \inf_{\tau\in\mathcal{N}} \sup_{\sigma\in\mathcal{C}} \Psi_i(\sigma,\tau) = g_i^\star. \tag{36}$$

For an initial distribution μ, the value is $\mu^\top g^\star$. The controller $\pi^\star$ from Theorem 3.2 is optimal for every stated criterion, state, and initial distribution.

The proof makes the simultaneous statewise guarantees explicit in the setting of Grand-Clement et al. (2023, Theorems 3.5–3.6 and Corollary 3.7). It combines the pathwise endpoints of Proposition D.2 with the common stationary strategies of Theorem E.1.

Proof of Theorem F.1. Fix Ψ, a state i, and $\delta > 0$. By equation 35, choose $q_\delta \in \mathcal{Q}_S$ with $d(q_\delta) \leq g^\star + \delta\mathbf{1}$.

Step 1: one-sided bounds. Proposition D.2 and $g^{\pi^\star} = g^\star$ give

$$I_i^-(\pi^\star,\tau) \geq g_i^\star \quad (\tau \in \mathcal{Q}_H), \qquad I_i^+(\sigma,q_\delta) \leq d_i(q_\delta) \leq g_i^\star + \delta \quad (\sigma \in \Pi_H).$$

By equation 25, the same bounds hold with Ψ in place of I^- and I^+.

Step 2: squeeze. Since $\pi^\star \in \mathcal{C}$ and $q_\delta \in \mathcal{N}$,

$$g_i^\star \leq \inf_{\tau\in\mathcal{N}} \Psi_i(\pi^\star,\tau) \leq \sup_{\sigma\in\mathcal{C}} \inf_{\tau\in\mathcal{N}} \Psi_i(\sigma,\tau) \leq \inf_{\tau\in\mathcal{N}} \sup_{\sigma\in\mathcal{C}} \Psi_i(\sigma,\tau) \leq \sup_{\sigma\in\mathcal{C}} \Psi_i(\sigma,q_\delta) \leq g_i^\star + \delta.$$

The middle inequality is weak duality. Letting $\delta \downarrow 0$ proves equation 36. The first two inequalities show that $\pi^\star$ attains the outer supremum.

Step 3: initial distributions. Let μ be an initial law, and write $I_\mu^\pm, J_\mu^\pm$ for the four payoffs when $S_0 \sim \mu$. The random variables $\liminf_N X_N$ and $\limsup_N X_N$ are bounded, and their expectations under μ are the μ-weighted sums of their conditional expectations given $S_0 = i$. Conditioning a strategy on $S_0 = i$ yields an admissible strategy from i. Step 1 therefore gives $I_\mu^-(\pi^\star,\tau) \geq \mu^\top g^\star$ and $I_\mu^+(\sigma,q_\delta) \leq \mu^\top g^\star + \delta$. The ordering equation 25 also holds under μ, so the squeeze of Step 2 identifies the value $\mu^\top g^\star$ for all four criteria. Only the pathwise endpoints are decomposed over initial states. No identity between the limit inferior of a mixture and a mixture of limits inferior is used. □

G Discounted characterizations of stationary policies

Theorem G.1. *For every $\pi \in \Pi_S$,*

$$\lim_{\epsilon \downarrow 0} \epsilon \|V_\epsilon - V_\epsilon^\pi\|_\infty = \|g^\star - g^\pi\|_\infty. \tag{37}$$

Consequently, π is average optimal from all states if and only if this limit is zero. Moreover, there exists $\epsilon_0 > 0$ such that

$$\pi \in \Pi_D, \qquad 0 < \epsilon < \epsilon_0, \qquad V_\epsilon^\pi = V_\epsilon \quad \Longrightarrow \quad g^\pi = g^\star. \tag{38}$$

The gap identity follows directly from the two discounted limits in Theorems 3.1–3.2. It records the exact limiting all-state loss, including for stationary randomized policies. The corresponding deterministic-policy statements are Grand-Clement et al. (2023, Theorems 4.6 and 4.10); the finiteness argument below gives equation 38 directly. The threshold is existential and supplies no computable stopping rule.

Proof of Theorem G.1. **Step 1: the gap identity.** By Lemma D.1(i) and Step 1 of the proof of Theorem E.1,

$$\epsilon(V_\epsilon - V_\epsilon^\pi) \longrightarrow g^\star - g^\pi \quad \text{in } \|\cdot\|_\infty.$$

Continuity of the norm gives equation 37. Lemma C.1(ii) gives $V_\epsilon^\pi \le V_\epsilon$, hence $g^\pi \le g^\star$. By Theorems 3.1 and F.1, the equality $g^\pi = g^\star$ holds exactly when π is average optimal from every state, for every payoff convention. This includes randomized stationary π.

Step 2: the discount threshold. Let $\mathcal{B} := \{\pi \in \Pi_D : g^\pi \neq g^\star\}$. If $\mathcal{B}$ is empty, take $\epsilon_0 := 1$. Otherwise, for each $\pi \in \mathcal{B}$, define its positive gain loss by $\Delta_\pi := \|g^\star - g^\pi\|_\infty > 0$. Step 1 gives a number $\epsilon_\pi > 0$ such that

$$\epsilon \|V_\epsilon - V_\epsilon^\pi\|_\infty > \Delta_\pi/2 \qquad (0 < \epsilon < \epsilon_\pi).$$

Thus this suboptimal controller cannot be discounted optimal at any such ϵ. Set $\epsilon_0 := \min_{\pi \in \mathcal{B}} \epsilon_\pi > 0$. Because $\mathcal{B}$ is finite, this one threshold excludes every average-suboptimal deterministic controller. This proves equation 38. □

H Exact nature attainment and stationary saddles

Recall that $d_i(q) = \max_{\pi \in \Pi_D} \eta_i^{\pi,q}$ is the optimal gain of the nominal MDP obtained by fixing the full selector $q \in \mathcal{Q}_S$ (Lemma C.6).

Theorem H.1. *The stationary performance vectors satisfy*

$$\begin{aligned} g_i^\star &= \max_{\pi \in \Pi_D} \inf_{q \in \mathcal{Q}_S} \eta_i^{\pi,q} = \inf_{q \in \mathcal{Q}_S} d_i(q), \qquad i \in S, \\ d(q) &\ge g^\star, \qquad q \in \mathcal{Q}_S. \end{aligned} \tag{39}$$

For every $\delta > 0$, one $q \in \mathcal{Q}_S$ satisfies $d(q) \le g^\star + \delta \mathbf{1}$.

A stationary selector q is exactly optimal for nature against every history-dependent controller, from every state and for every payoff in equation 1, if and only if

$$d(q) = g^\star. \tag{40}$$

For $\bar\pi \in \Pi_S$, $\bar q \in \mathcal{Q}_S$, and $u \in \mathbb{R}^S$, the stationary gain inequalities

$$\eta^{\bar\pi,q} \ge u \quad \forall q \in \mathcal{Q}_S, \qquad \eta^{\pi,\bar q} \le u \quad \forall \pi \in \Pi_D \tag{41}$$

hold if and only if $u = g^\star$ and $(\bar\pi, \bar q)$ is an all-state stationary saddle against history-dependent opponents for all four payoffs.

The dual representation builds on Grand-Clement et al. (2023, Theorem 3.5). The formulation through $d(q)$ isolates simultaneous exact attainment, while equation 41 expresses the full saddle property using only stationary-chain gains.

Proof of Theorem H.1. **Step 1: stationary duality.** The gain formula in Theorem E.1 and Lemma D.1(i) gives

$$g_i^\star = \max_{\pi\in\Pi_D} g_i^\pi = \max_{\pi\in\Pi_D} \inf_q \eta_i^{\pi,q}.$$

Step 3 of the same proof gives $d(q) \geq g^\star$ for every q, and equation 35 gives, for every $\delta > 0$, a selector with $d(q) \leq g^\star + \delta\mathbf{1}$. Hence $\inf_q d_i(q) = g_i^\star$ at every state, which proves equation 39 and the approximation claim.

Step 2: the exact attainment test. Suppose $d(q) = g^\star$. By Lemma C.6(iii), $\sup_{\sigma\in\Pi_H} \Psi_i(\sigma, q) = d_i(q) = g_i^\star$ for every payoff and state. Since $g_i^\star$ is the value (Theorem F.1), q is exactly optimal for nature. Conversely, suppose q secures $g^\star$ for at least one payoff Ψ, that is, $\Psi_i(\sigma, q) \leq g_i^\star$ for every $\sigma \in \Pi_H$ and i. Taking $\sigma = \pi \in \Pi_D$ and using $\Psi_i(\pi, q) = \eta_i^{\pi,q}$ (Lemma C.3) gives $d(q) \leq g^\star$. Together with Step 1, $d(q) = g^\star$. This proves equation 40, and it shows that exact attainment for one payoff convention already implies it for all four.

Step 3: the gain-only saddle test. Suppose equation 41 holds. Then

$$u \leq \inf_q \eta^{\bar\pi,q} = g^{\bar\pi} \leq g^\star \leq d(\bar q) = \max_{\pi\in\Pi_D} \eta^{\pi,\bar q} \leq u.$$

The equality on the left is equation 29, the inequality $g^{\bar\pi} \leq g^\star$ is Step 1 of the proof of Theorem G.1, and $g^\star \leq d(\bar q)$ is Step 1 above. All terms are therefore equal to $g^\star$. Proposition D.3 gives $\inf_{\tau\in\mathcal{Q}_H} \Psi_i(\bar\pi, \tau) = g_i^{\bar\pi} = g_i^\star$, and Lemma C.6(iii) gives $\sup_{\sigma\in\Pi_H} \Psi_i(\sigma, \bar q) = d_i(\bar q) = g_i^\star$. Hence $\Psi_i(\bar\pi, \tau) \geq g_i^\star \geq \Psi_i(\sigma, \bar q)$ for all $\sigma \in \Pi_H$, $\tau \in \mathcal{Q}_H$, and i, so $(\bar\pi, \bar q)$ is an all-state saddle at $g^\star$ for every payoff. Conversely, if $(\bar\pi, \bar q)$ is an all-state saddle at $g^\star$, restricting its guarantees to stationary opponents and using Lemma C.3 gives equation 41 with $u = g^\star$. □

A worst reply can leave a profitable deviation. Consider states s, L, H, with L, H absorbing and rewards $0, 0, 1$, respectively. At s, action a moves to L, and action b has row set $\{e_L, e_H\}$. Both deterministic policies satisfy $g^{\pi_a} = g^{\pi_b} = g^\star = (0, 0, 1)$, since nature answers b with e_L. Choose the full selector q with $q_{sb} = e_H$. Then $\eta^{\pi_a,q} = (0, 0, 1)$ but $\eta^{\pi_b,q} = (1, 0, 1)$. Thus q is exactly worst against the optimal controller π_a, yet $d(q) = (1, 0, 1) \neq g^\star$: the row at the unused action b permits a profitable deviation. The same example works with row set $\mathrm{co}\{e_L, e_H\}$.

I Proof of Theorem 4.1: vector Bellman verification

This section proves that every finite solution of the vector Bellman system gives the robust value and stationary optimal strategies. The first lemma explains why the gain and bias must be optimized in that order: it computes the limit of $T(tg + h) - tg$ as $t \to \infty$. The second lemma converts a bias inequality on gain-minimizing rows into an approximate inequality on every feasible row. The verification proof then sums these one-step inequalities along a trajectory. The resulting error is controlled by the expected change in the bounded quantity $g(S_t)$. We subsequently specialize the result to fixed policies and give two examples showing why both gain restrictions are necessary.

For a vector of candidate long-run gains g, the recession map retains only the continuation-gain terms in the Bellman operator:

$$(\widehat{T}g)_i := \max_{a\in A(i)} \min_{p\in\mathcal{U}_{ia}} p^\top g. \tag{42}$$

For every vector $g \in \mathbb{R}^S$, define the row minima and their minimizing sets by

$$m_{ia}(g) := \min_{p\in\mathcal{U}_{ia}} p^\top g, \qquad F_{ia}(g) := \{p \in \mathcal{U}_{ia} : p^\top g = m_{ia}(g)\}.$$

If $g = \widehat{T}g$, additionally put

$$A_g(i) := \{a \in A(i) : m_{ia}(g) = g_i\}, \qquad (L_g h)_i := \max_{a\in A_g(i)} \min_{p\in F_{ia}(g)} \{r_{ia} + p^\top h\}. \tag{43}$$

The continuous function $p \mapsto p^\top g$ attains its minimum on the compact set $\mathcal{U}_{ia}$, so $F_{ia}(g)$ is nonempty and compact. Finiteness of $A(i)$ and $g_i = \max_a m_{ia}(g)$ make $A_g(i)$ nonempty. We call $a \in A_g(i)$ a gain-active action and call $F_{ia}(g)$ a gain face, even when $\mathcal{U}_{ia}$ is nonconvex. We define $m_{ia}(g)$ and $F_{ia}(g)$ for arbitrary g because fixed-policy evaluation uses them when g does not solve the optimal-control gain equation.

A finite vector gain-bias certificate is a pair $(g, h) \in \mathbb{R}^S \times \mathbb{R}^S$ satisfying

$$g = \widehat{T}g, \qquad g + h = L_g h. \tag{44}$$

The first equation requires the candidate gain to be unchanged by gain optimization. The second equation compares the immediate reward and continuation bias only among actions and rows that optimize that first comparison. The face $F_{ia}(g)$ is defined relative to $m_{ia}(g)$ for every action, including inactive actions, for which $m_{ia}(g) < g_i$.

Lemma I.1. *If $g = \widehat{T}g$, then, for every $h \in \mathbb{R}^S$,*

$$T(tg + h) - tg \longrightarrow L_g h \quad \text{as } t \longrightarrow \infty. \tag{45}$$

The convergence is uniform when h ranges over any fixed compact subset of $\mathbb{R}^S$. Consequently, equation 44 is equivalent to

$$T(tg + h) = (t + 1)g + h + o(1) \qquad (t \to \infty). \tag{46}$$

If every ambiguity set is a polytope, then, for each fixed h, $T(tg + h) - tg = L_g h$ for all sufficiently large t.

Proof. **Step 1: the limit of one row.** Fix (i, a) and write $m := m_{ia}(g)$, $F := F_{ia}(g)$, and $b := \min_{p \in F}(r_{ia} + p^\top h)$. For $t \geq 0$ put

$$b_t := \min_{p \in \mathcal{U}_{ia}} \big\{ t(p^\top g - m) + r_{ia} + p^\top h \big\}.$$

We show $b_t \to b$. Every $p \in \mathcal{U}_{ia}$ has $p^\top g - m \geq 0$ and $p^\top h \geq -\|h\|_\infty$, so $b_t \geq r_{ia} - \|h\|_\infty$. A minimizer of $r_{ia} + p^\top h$ over F has zero gain gap, so $b_t \leq b$. Let p_t attain b_t. Then

$$0 \leq t(p_t^\top g - m) = b_t - r_{ia} - p_t^\top h \leq (b - r_{ia}) + \|h\|_\infty \leq 2\|h\|_\infty.$$

For $t > 0$, this implies $p_t^\top g - m \leq 2\|h\|_\infty / t$. Choose a sequence $t_k \to \infty$ along which b_{t_k} tends to its limit inferior. Compactness permits passage to a further subsequence, still denoted by t_k, such that $p_{t_k} \to p_0 \in \mathcal{U}_{ia}$. Continuity gives $p_0^\top g = m$, that is, $p_0 \in F$. Dropping the nonnegative gain term,

$$\liminf_{t \to \infty} b_t = \lim_k b_{t_k} \geq \lim_k (r_{ia} + p_{t_k}^\top h) = r_{ia} + p_0^\top h \geq b.$$

Together with $b_t \leq b$, this proves $b_t \to b$.

Step 2: the action maximum. Write $b_{ia,t}$ for the quantity b_t of Step 1 at the pair (i, a), and b_{ia} for its limit. Since $\min_p p^\top (tg + h) = t m_{ia}(g) + \min_p \{t(p^\top g - m_{ia}(g)) + p^\top h\}$,

$$[T(tg + h) - tg]_i = \max_{a \in A(i)} \big\{ t(m_{ia}(g) - g_i) + b_{ia,t} \big\}.$$

For an active action the term in braces tends to $b_{ia} = \min_{p \in F_{ia}(g)}(r_{ia} + p^\top h)$. For an inactive action, $m_{ia}(g) - g_i < 0$ and $b_{ia,t} \leq b_{ia}$, so the term tends to $-\infty$. Since $A(i)$ is finite and $A_g(i)$ is nonempty, the maximum tends to $\max_{a \in A_g(i)} b_{ia} = (L_g h)_i$. There are finitely many states, which proves equation 45 in supremum norm.

Step 3: uniformity on compact sets. The map $h \mapsto T(tg+h) - tg$ is nonexpansive by Lemma C.1(i). The map L_g is nonexpansive because each coordinate is a maximum of minima of affine maps $h \mapsto r_{ia} + p^\top h$ with stochastic p. Let K be compact, let $\delta > 0$, and let $h^1, \ldots, h^M$ be a finite δ-net of K. For $h \in K$ choose h^j with $\|h - h^j\|_\infty \leq \delta$; then

$$\|T(tg + h) - tg - L_g h\|_\infty \leq 2\delta + \max_{j \leq M} \|T(tg + h^j) - tg - L_g h^j\|_\infty.$$

Let $t \to \infty$, so that the finite maximum tends to zero, and then let $\delta \downarrow 0$.

Step 4: equivalence with the asymptotic ray. If equation 44 holds, then Steps 1–2 give $T(tg + h) - tg \to L_g h = g + h$, which is equation 46. Conversely, assume equation 46; this direction does not assume $g = \widehat{T}g$. Dividing by t gives $T(tg + h)/t \to g$. Each row objective in $T(tg + h)/t$ is $p^\top g + (r_{ia} + p^\top h)/t$, and the second term has absolute value at most $(R + \|h\|_\infty)/t$. Hence

$$\Big\| \frac{T(tg + h)}{t} - \widehat{T}g \Big\|_\infty \leq \frac{R + \|h\|_\infty}{t},$$

and therefore $g = \widehat{T}g$. Now L_g is defined, and Step 2 gives $L_g h = \lim_t \{T(tg + h) - tg\} = g + h$.

Step 5: exactness for polytopes. Suppose every $\mathcal{U}_{ia}$ is a polytope. A linear function on a polytope attains its minimum at a vertex, and $F_{ia}(g)$ is the face of $\mathcal{U}_{ia}$ spanned by the vertices it contains. Hence b_{ia} is the minimum of $r_{ia} + p^\top h$ over the vertices in $F_{ia}(g)$. A vertex $p \notin F_{ia}(g)$ has gap $d_p := p^\top g - m_{ia}(g) > 0$, and its objective in $b_{ia,t}$ is at least b_{ia} once

$$t \geq \frac{[b_{ia} - r_{ia} - p^\top h]_+}{d_p}.$$

Beyond the largest of these finitely many thresholds, $b_{ia,t} = b_{ia}$ for every pair. For an inactive action, the term $t(m_{ia}(g) - g_i) + b_{ia,t} \leq t(m_{ia}(g) - g_i) + b_{ia}$ falls below $(L_g h)_i$ after one more finite threshold, since $g_i - m_{ia}(g) > 0$. For all larger t, the maximum in Step 2 equals $(L_g h)_i$ exactly. □

Remark 8. *If $g = \rho\mathbf{1}$, then $p^\top g = \rho$ for every probability row. Therefore $A_g(i) = A(i)$, $F_{ia}(g) = \mathcal{U}_{ia}$, and $L_g = T$. The system becomes the scalar equation $\rho\mathbf{1} + h = T(h)$. The vector system thus extends scalar-gain Bellman equations without requiring a common gain across recurrent classes.*

I.1 A COMPACTNESS ESTIMATE FOR ROWS CLOSE TO A GAIN FACE

Lemma I.2. *Let X be compact, let $d, e : X \to \mathbb{R}$ be continuous, and suppose that $d \geq 0$ on X and $e \geq 0$ on $\{x : d(x) = 0\}$. For every $\theta > 0$ there is $C_\theta \in [0, \infty)$ such that*

$$e(x) \geq -\theta - C_\theta d(x) \qquad (x \in X). \tag{47}$$

If X is a polytope and d, e are affine, there is $C_0 < \infty$ for which the same inequality holds with $\theta = 0$.

Proof. **General case.** Fix $\theta > 0$; if X is empty there is nothing to prove. Put $M := \max_{x \in X}[-e(x)]_+$, which is finite by compactness, and $B_\theta := \{x \in X : e(x) \leq -\theta\}$. If B_θ is empty, then $e > -\theta$ on X, and choosing $C_\theta := 0$ works. Otherwise B_θ is a nonempty compact set. It contains no zero of d, because $e \geq 0$ wherever $d = 0$. Since d is continuous and positive on B_θ, the number $\delta_\theta := \min_{x \in B_\theta} d(x)$ is positive. Put $C_\theta := M/\delta_\theta$. On B_θ,

$$C_\theta d(x) \geq C_\theta \delta_\theta = M \geq -e(x),$$

so $e(x) \geq -C_\theta d(x) \geq -\theta - C_\theta d(x)$. Outside B_θ, $e(x) > -\theta \geq -\theta - C_\theta d(x)$ because $d \geq 0$.

Polytopes. Let $\mathcal{V}$ be the vertex set of X and put

$$C_0 := \max_{v \in \mathcal{V}:\, d(v) > 0} \frac{[-e(v)]_+}{d(v)},$$

with an empty maximum equal to zero. At a vertex with $d(v) > 0$, $e(v) + C_0 d(v) \geq e(v) + [-e(v)]_+ \geq 0$. At a vertex with $d(v) = 0$, the hypothesis gives $e(v) \geq 0$. The affine function $e + C_0 d$ is therefore nonnegative at every vertex, hence on their convex hull X. □

The additive θ cannot be removed in general. On $X := [0, 1]$, the functions $d(x) := x^2$ and $e(x) := -x$ satisfy the hypotheses, but no finite C_0 satisfies $-x \geq -C_0 x^2$ for all small $x > 0$. This distinction matters when the estimate is summed over N stages: compactness gives an error that is $o(N)$, but that error need not be bounded independently of N.

I.2 VERIFICATION AGAINST HISTORY-DEPENDENT OPPONENTS

Nature's stationary strategy must specify a row for every state and every controller action. Rows specified only for the actions used by the selected controller would not define a strategy against a different controller.

Theorem I.3. *Suppose equation 44 has a finite solution (g, h). At every state choose*

$$\pi^*(i) \in \operatorname*{argmax}_{a \in A_g(i)} \min_{p \in F_{ia}(g)} \{r_{ia} + p^\top h\}, \tag{48}$$

$$q^*_{ia} \in \begin{cases} \operatorname{argmin}_{p \in F_{ia}(g)} p^\top h, & a \in A_g(i), \\ F_{ia}(g), & a \notin A_g(i), \end{cases} \tag{49}$$

which specify a deterministic stationary policy and a stationary kernel. Then there exist a constant $C < \infty$ and, for every $\theta > 0$, a constant $C_\theta < \infty$ such that, for every initial state i, every horizon $N \geq 1$, and all randomized history-dependent strategies σ, τ,

$$\mathbb{E}_i^{\pi^*,\tau} \sum_{t=0}^{N-1} r_{S_t A_t} \geq N(g_i - \theta) - \operatorname{sp}(h) - C_\theta \operatorname{sp}(g), \tag{50}$$

$$\mathbb{E}_i^{\sigma,q^*} \sum_{t=0}^{N-1} r_{S_t A_t} \leq N g_i + \operatorname{sp}(h) + C \operatorname{sp}(g). \tag{51}$$

The constants can be chosen uniformly over all selectors satisfying equation 48–equation 49, with dependence

$$C = C(r, \mathcal{U}, g, h) < \infty, \qquad C_\theta = C_\theta(r, \mathcal{U}, g, h, \theta) < \infty. \tag{52}$$

They do not depend on the initial state, the horizon, or the opposing strategies. In general, C_θ need not remain bounded as $\theta \downarrow 0$; see the paragraph after the proof. In particular,

$$\liminf_{N\to\infty} J_N(i; \pi^*, \tau) \geq g_i, \qquad \limsup_{N\to\infty} J_N(i; \sigma, q^*) \leq g_i. \tag{53}$$

For each $\Psi \in \{I^-, J^-, J^+, I^+\}$, both the max-min and the min-max values equal g_i, and (π^*, q^*) is an all-state stationary saddle against history-dependent opponents. These equalities remain valid when either or both strategy classes are restricted to stationary strategies. Moreover,

$$\mathbb{E}_i^{\pi^*,q^*} \sum_{t=0}^{N-1} r_{S_t A_t} = N g_i + h_i - \mathbb{E}_i^{\pi^*,q^*} h(S_N), \tag{54}$$

and

$$\frac{T^N 0}{N} \longrightarrow g. \tag{55}$$

Thus the gain component is the same for every finite certificate, and it equals $g^\star$. If the ambiguity sets of the actions $\pi^*(i)$ are polytopes, then equation 50 also holds with $\theta = 0$ and a finite C_0.

Proof. Fix the candidate certificate (g, h). The proof first establishes one-step inequalities, then sums them to obtain the expected-payoff bounds. A separate martingale argument gives the payoffs with a sample-path limit inside the expectation. We use the filtrations from Appendix C: $\mathcal{F}_t$ records the history before the stage-t action and row are drawn, while $\mathcal{G}_t$ additionally records that action and row. In particular, $\mathcal{F}_t \subseteq \mathcal{G}_t \subseteq \mathcal{F}_{t+1}$ and $\mathbb{E}[f(S_{t+1}) \mid \mathcal{G}_t] = p_t^\top f$.

Part 1: controller inequalities on all feasible rows. Define

$$A^\star_{g,h}(i) := \operatorname*{argmax}_{a \in A_g(i)} \min_{p \in F_{ia}(g)} \{r_{ia} + p^\top h\}.$$

For $a \in A^\star_{g,h}(i)$ and $p \in \mathcal{U}_{ia}$, define the gain drift and bias residual by

$$d_{ia}(p) := p^\top g - g_i, \qquad e_{ia}(p) := r_{ia} + p^\top h - g_i - h_i.$$

Since a is gain active, $d_{ia} \geq 0$ on $\mathcal{U}_{ia}$, and its zero set is $F_{ia}(g)$. Since a attains the maximum defining $(L_g h)_i = g_i + h_i$, we have $\min_{p \in F_{ia}(g)} e_{ia}(p) = 0$, so $e_{ia} \geq 0$ on the zero set of d_{ia}. Lemma I.2 with $X = \mathcal{U}_{ia}$ gives a constant for each such pair. Let C_θ be the maximum of these constants over the finitely many pairs (i, a) with $a \in A^\star_{g,h}(i)$. For every selector satisfying equation 48,

$$e_{i,\pi^*(i)}(p) \geq -\theta - C_\theta d_{i,\pi^*(i)}(p) \qquad (p \in \mathcal{U}_{i,\pi^*(i)}). \tag{56}$$

This constant does not depend on the tie-breaking in equation 48.

Part 2: the controller's finite-horizon guarantee. Fix a nature strategy τ and let the controller use π^*. Write $d_t := d_{S_t A_t}(p_t)$ and $e_t := e_{S_t A_t}(p_t)$, where $A_t = \pi^*(S_t)$. By equation 17,

$$\mathbb{E}[g(S_{t+1}) - g(S_t) \mid \mathcal{G}_t] = d_t \geq 0.$$

Conditioning once more on $\mathcal{F}_t$ shows that $g(S_t)$ is a submartingale: its conditional expected next value is at least its current value. In particular, $\mathbb{E}_i g(S_t) \geq g_i$. Taking expectations and summing,

$$\sum_{t=0}^{N-1} \mathbb{E}_i d_t = \mathbb{E}_i g(S_N) - g_i \leq \operatorname{sp}(g). \tag{57}$$

The definition of e_t gives $r_{S_tA_t} = g(S_t) + h(S_t) - p_t^\top h + e_t$. Since $p_t^\top h = \mathbb{E}[h(S_{t+1}) \mid \mathcal{G}_t]$, taking expectations and summing yields the exact expected reward identity

$$\mathbb{E}_i \sum_{t=0}^{N-1} r_{S_tA_t} = \sum_{t=0}^{N-1} \mathbb{E}_i g(S_t) + h_i - \mathbb{E}_i h(S_N) + \sum_{t=0}^{N-1} \mathbb{E}_i e_t. \tag{58}$$

The intermediate bias terms cancel between consecutive stages. Now insert the bounds $\mathbb{E}_i g(S_t) \geq g_i$, $h_i - \mathbb{E}_i h(S_N) \geq -\operatorname{sp}(h)$, $\mathbb{E}_i e_t \geq -\theta - C_\theta \mathbb{E}_i d_t$ from equation 56, and equation 57. This gives equation 50. Dividing by N, taking the limit inferior at fixed θ, and then letting $\theta \downarrow 0$ proves the controller half of equation 53; neither the selector nor the opponent depends on θ.

Part 3: nature's full selector and upper guarantee. For every state-action pair, including actions not used by π^*, define

$$\bar{d}_{ia} := (q_{ia}^*)^\top g - g_i, \qquad \bar{e}_{ia} := r_{ia} + (q_{ia}^*)^\top h - g_i - h_i.$$

For an active action, $q_{ia}^* \in F_{ia}(g)$ gives $\bar{d}_{ia} = m_{ia}(g) - g_i = 0$. Since q_{ia}^* minimizes $p^\top h$ over $F_{ia}(g)$ and $(L_g h)_i$ is the maximum over active actions,

$$\bar{e}_{ia} = \min_{p \in F_{ia}(g)} (r_{ia} + p^\top h) - (L_g h)_i \leq 0.$$

For an inactive action, $q_{ia}^* \in F_{ia}(g)$ gives $\bar{d}_{ia} = m_{ia}(g) - g_i < 0$. Its bias residual may be positive. Because its gain drift is strictly negative and there are finitely many actions, a finite constant bounds this residual in terms of the magnitude of that drift. Set

$$C := \max_{(i,a):\, a \notin A_g(i)} \frac{\max_{p \in F_{ia}(g)} [r_{ia} + p^\top h - g_i - h_i]_+}{g_i - m_{ia}(g)},$$

with an empty maximum equal to zero. Compactness bounds the numerators, and there are finitely many positive denominators. Thus C is finite and does not depend on nature's tie-breaking, and

$$\bar{d}_{ia} \leq 0, \qquad \bar{e}_{ia} \leq C(-\bar{d}_{ia}) \quad \text{for every } (i,a). \tag{59}$$

Now let the controller use any randomized history-dependent σ, and put $\bar{d}_t := \bar{d}_{S_tA_t}$ and $\bar{e}_t := \bar{e}_{S_tA_t}$. The inequalities equation 59 hold at the realized action, so $\mathbb{E}[g(S_{t+1}) - g(S_t) \mid \mathcal{G}_t] = \bar{d}_t \leq 0$. After conditioning on $\mathcal{F}_t$, this says that $g(S_t)$ is a supermartingale: its conditional expected next value is at most its current value. Consequently,

$$\mathbb{E}_i g(S_t) \leq g_i, \qquad \sum_{t<N} \mathbb{E}_i(-\bar{d}_t) = g_i - \mathbb{E}_i g(S_N) \leq \operatorname{sp}(g).$$

The identity equation 58, with $\bar{e}_t$ in place of e_t, now gives

$$\mathbb{E}_i^{\sigma,q^*} \sum_{t<N} r_{S_tA_t} \leq N g_i + \operatorname{sp}(h) + C \operatorname{sp}(g),$$

which is equation 51. Dividing by N gives the nature half of equation 53.

Part 4: the two pathwise payoff conventions. Bounds on expected averages do not by themselves bound $\mathbb{E} \liminf$ or $\mathbb{E} \limsup$, so we argue pathwise. For either one-sided strategy pair define

$$\xi_{t+1} := h(S_{t+1}) - p_t^\top h, \qquad M_N := \sum_{t=0}^{N-1} \xi_{t+1}.$$

By the transition rule and the tower property, $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0$. Thus M_N is a martingale with respect to $(\mathcal{F}_N)$, and its increments have absolute value at most $\operatorname{sp}(h)$. These fluctuations have a vanishing average: $M_N/N \to 0$ almost surely. Indeed, if $\operatorname{sp}(h) > 0$, the Azuma–Hoeffding inequality gives, for each $\varepsilon > 0$,

$$\mathbb{P}_i(|M_N| \geq N\varepsilon) \leq 2 \exp\left(-\frac{N\varepsilon^2}{2 \operatorname{sp}(h)^2}\right).$$

The right-hand side is summable in N, so the Borel–Cantelli lemma, applied to $\varepsilon = 1, 1/2, 1/3, \ldots$, proves the assertion. If $\operatorname{sp}(h) = 0$, then $M_N = 0$ identically. Substituting $p_t^\top h = h(S_{t+1}) - \xi_{t+1}$ into the definition of e_t gives the sample-path identity

$$\sum_{t<N} r_{S_tA_t} = \sum_{t<N} g(S_t) + h_i - h(S_N) + \sum_{t<N} e_t + M_N,$$

with $\bar{e}_t$ in place of e_t for the nature pair.

Under (π^*, τ), the martingale convergence theorem gives $g(S_t) \to G_\infty$ almost surely. The bound $|g(S_t)| \le \|g\|_\infty$ also gives convergence in L^1, so $\mathbb{E}_i G_\infty = \lim_t \mathbb{E}_i g(S_t) \ge g_i$. Monotone convergence and equation 57 give $\mathbb{E}_i \sum_{t\ge 0} d_t \le \mathrm{sp}(g)$, so $\sum_{t\ge 0} d_t < \infty$ almost surely. Fix $\theta > 0$. By equation 56, $e_t \ge -\theta - C_\theta d_t$. Dividing the path identity by N, the averages of $g(S_t)$ tend to G_∞, the average of $C_\theta d_t$ tends to zero, and $(h_i - h(S_N) + M_N)/N \to 0$. Hence

$$\liminf_N \frac{1}{N} \sum_{t<N} r_{S_t A_t} \ge G_\infty - \theta \quad \text{almost surely.}$$

Intersecting the probability-one events for $\theta = 1, 1/2, 1/3, \ldots$ gives the same bound with $\theta = 0$. Taking expectations, $I_i^-(\pi^*, \tau) \ge \mathbb{E}_i G_\infty \ge g_i$.

Under (σ, q^*), the same convergence theorem and boundedness give $g(S_t) \to \bar{G}_\infty$ almost surely and in L^1, with $\mathbb{E}_i \bar{G}_\infty \le g_i$. The nonnegative cumulative drift $\sum_t (-\bar{d}_t)$ has expectation at most $\mathrm{sp}(g)$ by Part 3, so it is finite almost surely. By equation 59 and the path identity,

$$\limsup_N \frac{1}{N} \sum_{t<N} r_{S_t A_t} \le \bar{G}_\infty \quad \text{almost surely,}$$

so $I_i^+(\sigma, q^*) \le g_i$. By equation 25, these two guarantees bracket every payoff Ψ, and weak duality gives

$$g_i \le \sup_\sigma \inf_\tau \Psi_i(\sigma, \tau) \le \inf_\tau \sup_\sigma \Psi_i(\sigma, \tau) \le g_i.$$

The same argument holds for any restricted strategy classes that contain π^* and q^*, including the stated stationary classes.

Part 5: the selected pair, the finite-horizon limit, and uniqueness. When both selected strategies are used, the gain and bias residuals vanish:

$$P^{\pi^*, q^*} g = g, \qquad r^{\pi^*} + P^{\pi^*, q^*} h = g + h.$$

The first identity holds because $q^*_{i,\pi^*(i)} \in F_{i,\pi^*(i)}(g)$ and $\pi^*(i)$ is active. The second holds because $\pi^*(i)$ attains the maximum in $(L_g h)_i$ and $q^*_{i,\pi^*(i)}$ attains the corresponding minimum. Under the selected pair, $e_t = 0$, and $P^{\pi^*, q^*} g = g$ gives $\mathbb{E}_i g(S_t) = g_i$ for every t. The identity equation 58 is therefore exactly equation 54. By Lemma C.1(iv), $T^N 0$ is the finite-horizon value, so the two uniform guarantees give

$$g_i - \theta - \frac{\mathrm{sp}(h) + C_\theta \,\mathrm{sp}(g)}{N} \le \frac{(T^N 0)_i}{N} \le g_i + \frac{\mathrm{sp}(h) + C \,\mathrm{sp}(g)}{N}.$$

Let $N \to \infty$ at fixed θ, and then let $\theta \downarrow 0$. Since there are finitely many states, this proves equation 55 in norm. The limit of $T^N 0/N$ does not depend on the certificate, and it equals $g^\star$ by Theorem 3.2; hence every finite certificate has gain $g^\star$. Finally, suppose the row set of each selected action $\pi^*(i)$ is a polytope. Apply the affine part of Lemma I.2 to these finitely many selected row sets and take the largest resulting constant. Repeating Part 2 with this $C_0 < \infty$ and $\theta = 0$ proves the last assertion. This final assertion only needs the selected-action row sets to be polytopes. □

Why C_θ can diverge as $\theta \downarrow 0$. If a sequence $\theta_n \downarrow 0$ admitted constants bounded by some $\bar{C} < \infty$, then, for each fixed N, letting $n \to \infty$ in equation 50 would give

$$(T^N 0)_i \ge N g_i - \mathrm{sp}(h) - \bar{C} \,\mathrm{sp}(g).$$

Together with equation 51, this would imply an $O(1/N)$ error for $T^N 0/N$. Proposition L.3 constructs a model with a finite certificate whose gain-estimation error instead has order $N^{-1/(k-1)}$, where $k \ge 3$. Therefore, in that model, every choice of constants satisfying the finite-horizon bound must have $C_\theta \to \infty$ as $\theta \downarrow 0$.

Remark 9. *Theorem I.3 proves uniqueness of the gain by identifying it with* $\lim_N T^N 0/N$. *The bias has a different status. Adding any constant multiple of* $\mathbf{1}$ *preserves its equation. More generally, put*

$$\mathcal{E}_g := \{d \in \mathbb{R}^S : p^\top d = d_i \text{ for all } i,\ a \in A_g(i),\ p \in F_{ia}(g)\}.$$

For $d \in \mathcal{E}_g$, every expression in the active bias optimization changes by exactly d_i. Hence $L_g(h+d) = L_g h + d$, so $h + d$ is another bias whenever h is. The space $\mathcal{E}_g$ contains both $\mathbf{1}$ and g. For an MDP consisting of absorbing states, $g_i = r_i$ and every vector h solves the bias equation. Thus a single reference-state normalization cannot in general determine a multichain bias, and even classwise normalizations require an additional uniqueness argument in a particular model.

Remark 10. *The selected controller and nature strategy do not depend on the initial state, the horizon, or the opponent. The constants in equation 50–equation 51 do not depend on these quantities either. Thus, for any accuracy, one horizon threshold makes the expected-average guarantees valid against all opponents simultaneously. These are statements about expected payoffs. They do not assert that every trajectory against every opponent has average reward at least, or at most, the deterministic number g_i.*

Corollary I.4. *Under (π^*, q^*) there is a bounded random variable G_∞, the gain of the recurrent class eventually entered, such that*

$$g(S_t) \longrightarrow G_\infty \quad \textit{almost surely}, \qquad \frac{1}{N}\sum_{t=0}^{N-1} r_{S_t A_t} \longrightarrow G_\infty \quad \textit{almost surely}, \qquad \mathbb{E}_i G_\infty = g_i. \tag{60}$$

In every recurrent class of the selected finite Markov chain, g is constant and equals the average reward of that class. From a transient state, g_i is the average of these class gains weighted by the absorption probabilities.

Proof. Write $P^* := P^{\pi^*, q^*}$ and $r_i^* := r_{i,\pi^*(i)}$. Part 5 of the preceding proof gives

$$P^* g = g, \qquad (I - P^*)h = r^* - g. \tag{61}$$

Let $(P^*)^\infty$ denote the Cesàro projector of Lemma C.2, that is, the limit of $N^{-1}\sum_{t=0}^{N-1}(P^*)^t$. Since $P^* g = g$, every term $(P^*)^t g$ equals g, and therefore $(P^*)^\infty g = g$. Apply the projector to the second equation in equation 61. The identity $(P^*)^\infty(I - P^*) = 0$ then gives $(P^*)^\infty r^* = g$.

For a finite Markov chain, this last vector records the average reward in each recurrent class and its absorption-probability average at transient states, as stated in Lemma C.3. The chain almost surely enters a recurrent class, and the sample reward average converges to that class's reward. The equation $P^* g = g$ makes g constant on each recurrent class, with that same value. Calling this random class value G_∞ gives both almost-sure limits. Bounded convergence and $\mathbb{E}_i g(S_t) = g_i$ give $\mathbb{E}_i G_\infty = g_i$. Periodicity does not affect these average-reward conclusions. □

I.3 Fixed-policy specializations of the verification theorem

For a deterministic stationary controller π, put

$$(T^\pi x)_i := r_{i,\pi(i)} + \min_{p \in \mathcal{U}_{i,\pi(i)}} p^\top x, \qquad (\widehat{T}^\pi g)_i := \min_{p \in \mathcal{U}_{i,\pi(i)}} p^\top g.$$

Given $g = \widehat{T}^\pi g$, define $F_i^\pi(g) := \{p \in \mathcal{U}_{i,\pi(i)} : p^\top g = g_i\}$. The two evaluation equations are

$$g = \widehat{T}^\pi g, \qquad g_i + h_i = r_{i,\pi(i)} + \min_{p \in F_i^\pi(g)} p^\top h \quad (i \in S). \tag{62}$$

The first equation identifies the rows that minimize continuation gain. The second minimizes the bias over those rows. Replacing the scalar gain by a vector in the single equation $g + h = T^\pi h$ does not enforce this restriction.

Proposition I.5. *If (g, h) solves equation 62, choose*

$$q_i^* \in \operatorname*{argmin}_{p \in F_i^\pi(g)} p^\top h.$$

Then, for every initial state,

$$\inf_\tau \liminf_N J_N(i;\pi,\tau) = \inf_\tau \limsup_N J_N(i;\pi,\tau) = g_i, \qquad \frac{(T^\pi)^N 0}{N} \longrightarrow g. \tag{63}$$

The infima may be taken over all history-dependent randomized nature strategies or over stationary strategies only. Extend the selected rows to a full nature strategy by setting $q^*_{i,\pi(i)} := q^*_i$ and choosing arbitrary feasible rows at unused actions. This single stationary strategy attains both infima at every initial state. Under (π, q^*), the identity equation 54 and the pathwise conclusions of Corollary I.4 hold.

Proof. Consider the model in which the controller has only the action $\pi(i)$ at state i. In this one-action model, $A_g(i) = \{\pi(i)\}$ by the first equation in equation 62, $F_{i\pi(i)}(g) = F^\pi_i(g)$, and the gain and bias equations equation 44 are exactly equation 62. Theorem I.3 therefore applies, with π as the only controller and q^* as nature's selector. For every $\theta > 0$ and every history-dependent τ,

$$J_N(i;\pi,\tau) \geq g_i - \theta - \frac{\operatorname{sp}(h) + C_\theta \operatorname{sp}(g)}{N}, \qquad J_N(i;\pi,q^*) = g_i + \frac{h_i - \mathbb{E}^{\pi,q^*}_i h(S_N)}{N}.$$

Since every value of $h(S_N)$ lies between the smallest and largest coordinates of h, the numerator in the second expression has absolute value at most $\operatorname{sp}(h)$. The first bound gives $\inf_\tau \liminf_N J_N \geq g_i$, and the second gives $\inf_\tau \limsup_N J_N \leq \lim_N J_N(i;\pi,q^*) = g_i$. Since the limit inferior never exceeds the limit superior, both infima equal g_i and are attained by q^*. The finite-horizon limit and the pathwise conclusions are the corresponding conclusions of Theorem I.3 and Corollary I.4 for the one-action model. □

I.3.1 Randomized stationary policies and action-contingent nature

Let $\pi(a \mid i)$ be a fixed stationary randomized policy, and recall that nature observes the realized action. As in equation 20, the expected one-step reward and the set of effective transition rows are

$$r^\pi_i := \sum_{a \in A(i)} \pi(a \mid i) r_{ia}, \qquad \mathcal{U}^\pi_i := \left\{ \sum_{a \in A(i)} \pi(a \mid i) p_a : p_a \in \mathcal{U}_{ia} \text{ for all } a \right\}. \tag{64}$$

The set $\mathcal{U}^\pi_i$ is nonempty and compact. Actions of probability zero can be omitted from the product without changing $\mathcal{U}^\pi_i$.

Proposition I.6. *For the post-action nature model, the fixed-policy Bellman operator is*

$$(T^\pi x)_i := \sum_a \pi(a \mid i) \left(r_{ia} + \min_{p \in \mathcal{U}_{ia}} p^\top x \right) = r^\pi_i + \min_{\bar{p} \in \mathcal{U}^\pi_i} \bar{p}^\top x. \tag{65}$$

Its finite vector evaluation certificate is

$$g_i = \sum_a \pi(a \mid i) m_{ia}(g), \tag{66}$$

$$g_i + h_i = \sum_a \pi(a \mid i) \left(r_{ia} + \min_{p \in F_{ia}(g)} p^\top h \right). \tag{67}$$

Every finite solution has all the expected-value conclusions of Proposition I.5. A worst stationary nature strategy is obtained by choosing, separately for every positive-probability action,

$$q^*_{ia} \in \operatorname*{argmin}_{p \in F_{ia}(g)} p^\top h. \tag{68}$$

In this statement g may be nonconstant, and the individual quantities $m_{ia}(g)$ need not equal g_i.

Proof. **Step 1: the fixed-policy operator.** At state i, action a occurs with probability $\pi(a \mid i)$. Nature observes that action and can minimize its continuation value separately. Thus the expected immediate reward is r^π_i, and the smallest expected continuation value is $\sum_a \pi(a \mid i) \min_{p \in \mathcal{U}_{ia}} p^\top x$. This proves the first expression in equation 65. Minimizing a weighted sum over independent row choices gives the second expression. Applying the same identity to g shows that equation 66 is precisely the gain equation for the one-action model with row sets $\mathcal{U}^\pi_i$.

Step 2: rows that minimize the effective gain. Assume equation 66. For any tuple of feasible rows $(p_a)_a$, set $\bar{p} := \sum_a \pi(a \mid i) p_a$. Then

$$\bar{p}^\top g - g_i = \sum_a \pi(a \mid i) \big(p_a^\top g - m_{ia}(g) \big).$$

Every term in parentheses is nonnegative. Consequently, $\bar{p}^\top g = g_i$ if and only if $p_a \in F_{ia}(g)$ for every action with positive probability. This conclusion holds for every representation of $\bar{p}$. Therefore the effective gain face consists exactly of the weighted sums of component gain faces, and

$$\min_{\bar{p}\in\mathcal{U}_i^\pi:\,\bar{p}^\top g=g_i} \bar{p}^\top h = \sum_a \pi(a \mid i) \min_{p\in F_{ia}(g)} p^\top h.$$

Equation equation 67 is thus the bias equation of the effective one-action model. The choices in equation 68 attain its minimum. Complete q^* by assigning arbitrary feasible rows to actions of probability zero.

Step 3: the lower bound against adaptive, randomized nature. We must also justify using the effective-row calculation when nature chooses its row after seeing the action. Fix any such strategy τ. For every state i, apply Lemma I.2 on $\mathcal{U}_i^\pi$ to

$$d_i(\bar{p}) := \bar{p}^\top g - g_i, \qquad e_i(\bar{p}) := r_i^\pi + \bar{p}^\top h - g_i - h_i.$$

Steps 1–2 give $d_i \geq 0$ and $e_i \geq 0$ wherever $d_i = 0$. Taking the largest constant over the finitely many states gives

$$e_i(\bar{p}) \geq -\theta - C_\theta d_i(\bar{p}) \qquad (\bar{p} \in \mathcal{U}_i^\pi).$$

Condition now on $\mathcal{F}_t$, before the action is drawn. Let $\kappa_{t,a}$ denote the conditional law of nature's row given this history and the event $A_t = a$. If that action has probability zero, choose any law on $\mathcal{U}_{S_t a}$. Define

$$\bar{p}_t := \sum_a \pi(a \mid S_t) \int p\, \kappa_{t,a}(dp), \qquad d_t := \bar{p}_t^\top g - g(S_t), \qquad e_t := r_{S_t}^\pi + \bar{p}_t^\top h - g(S_t) - h(S_t).$$

The preceding inequality holds for every tuple $(p_a)_a$ before averaging. Integrating it over the product of the laws $\kappa_{t,a}$ gives $d_t \geq 0$ and $e_t \geq -\theta - C_\theta d_t$. This step does not require $\bar{p}_t \in \mathcal{U}_{S_t}^\pi$, which can fail when the row sets are nonconvex.

The transition rule and the fixed action probabilities give

$$\mathbb{E}[f(S_{t+1}) \mid \mathcal{F}_t] = \bar{p}_t^\top f, \qquad \mathbb{E}[r_{S_t A_t} \mid \mathcal{F}_t] = r_{S_t}^\pi.$$

In particular, d_t is the conditional expected increment of $g(S_t)$. Therefore $\mathbb{E}_i g(S_t) \geq g_i$ and $\sum_{t<N} \mathbb{E}_i d_t = \mathbb{E}_i g(S_N) - g_i \leq \operatorname{sp}(g)$. Taking expectations in the definition of e_t and summing yields

$$\mathbb{E}_i^{\pi,\tau} \sum_{t<N} r_{S_t A_t} = \sum_{t<N} \mathbb{E}_i g(S_t) + h_i - \mathbb{E}_i h(S_N) + \sum_{t<N} \mathbb{E}_i e_t \geq N(g_i - \theta) - \operatorname{sp}(h) - C_\theta \operatorname{sp}(g).$$

This is the required lower bound for every history-dependent nature strategy.

Step 4: attainment and finite-horizon convergence. For the stationary pair (π, q^*), Step 2 gives

$$P^{\pi,q^*} g = g, \qquad r^\pi + P^{\pi,q^*} h = g + h.$$

Iterating the first equality preserves the expected gain, while summing the second makes the intermediate bias terms cancel. Hence

$$\mathbb{E}_i^{\pi,q^*} \sum_{t<N} r_{S_t A_t} = N g_i + h_i - \mathbb{E}_i^{\pi,q^*} h(S_N).$$

The last two terms have absolute difference at most $\operatorname{sp}(h)$, so $J_N(i;\pi,q^*) \to g_i$. Combined with the lower bound of Step 3, this proves both equalities in equation 63, with the same stationary q^* attaining them from every state. Finally, $[(T^\pi)^N 0]_i$ is nature's finite-horizon value against π by Lemma C.1(iv). It is bounded below by Step 3 and above by the payoff of q^*. Dividing by N, then letting $N \to \infty$ and $\theta \downarrow 0$, proves $(T^\pi)^N 0/N \to g$. □

Remark 11 (Why one must average before using the randomized-policy gain)**.** *For a randomized fixed policy, equation 66 equates only the weighted average of the $m_{ia}(g)$ with g_i. It need not imply $p^\top g \geq g_i$ separately for every realized action. Accordingly, the submartingale argument conditions on the history before the action is sampled, or equivalently uses the effective-row model. Applying the deterministic-policy argument to each realized action separately would be incorrect.*

I.4 WHY BOTH GAIN RESTRICTIONS ARE NECESSARY

The following two counterexamples show separately that nature's minimizing face and the controller's active action set are necessary. They use the same transition geometry, so the controller example reuses the absorption calculation from the nature example.

Example 1 (Nature's gain restriction cannot be omitted)**.** *There are states (s, m, z) with rewards $(0, -1, 0)$, and m, z are absorbing. At s, let*

$$p^\lambda := (1 - \lambda/2, \lambda/4, \lambda/4), \qquad \mathcal{U}_s := \{p^\lambda : 0 \le \lambda \le 1\}.$$

The robust gain is $g := (-1/2, -1, 0)$, and $h := (0, -2, 0)$ is a Bellman bias. However, the unrestricted equations $\widehat{T}\widetilde{g} = \widetilde{g}$ and $T\widetilde{h} = \widetilde{g} + \widetilde{h}$ also accept the incorrect pair $\widetilde{g} := (-3/4, -1, 0)$, $\widetilde{h} := (0, -3, 0)$.

Proof. **The robust gain.** Define the hitting times $\tau_m := \inf\{t \ge 0 : S_t = m\}$ and $\tau_z := \inf\{t \ge 0 : S_t = z\}$, with $\inf \varnothing := \infty$. Let $\lambda_t \in [0, 1]$ denote nature's parameter at time t, allowing dependence on the past and randomization. The two first-entry probabilities at time $t + 1$ are equal:

$$\mathbb{P}_s(\tau_m = t + 1) = \mathbb{E}_s[\mathbf{1}_{\{S_t = s\}} \lambda_t / 4] = \mathbb{P}_s(\tau_z = t + 1).$$

Summing over t shows that each eventual absorption probability is at most $1/2$. On a trajectory that reaches m, the sample reward average tends to -1. On every other trajectory it equals zero throughout. Thus the sample average converges to $-\mathbf{1}_{\{\tau_m < \infty\}}$. Bounded convergence shows that all four payoff conventions equal $-\mathbb{P}_s(\tau_m < \infty) \ge -1/2$. Taking $\lambda_t = 1$ at every visit to s attains $-1/2$, because the probability of remaining at s for t steps is 2^{-t} and the two absorbing states are then reached with probability $1/2$ each.

The certificate. For every λ,

$$(p^\lambda)^\top g = -1/2, \qquad (p^\lambda)^\top h = -\lambda/2.$$

Every row is therefore gain minimizing, and the gain-face bias equation at s reads $-1/2 + 0 = 0 + \min_\lambda(-\lambda/2)$, which holds. The equations at the absorbing states are identities.

The incorrect pair. Direct substitution gives

$$(p^\lambda)^\top \widetilde{g} = -3/4 + \lambda/8, \qquad (p^\lambda)^\top \widetilde{h} = -3\lambda/4.$$

The gain minimum is $-3/4$, attained only at $\lambda = 0$, whereas the unrestricted bias minimum is $-3/4$, attained at $\lambda = 1$. The unrestricted equations combine these incompatible choices. On the actual gain face $\{p^0\}$, the bias equation would require $-3/4 + 0 = 0 + 0$, which is false. □

Example 2 (The controller's gain restriction cannot be omitted)**.** *Use states (s, m, z) with rewards $(0, 1, 0)$, and make m, z absorbing. At s, the controller chooses between the nominal rows $p^0 := (1, 0, 0)$ and $p^1 := (1/2, 1/4, 1/4)$. The true gain is $(1/2, 1, 0)$, but the unrestricted equations accept $\widetilde{g} := (3/4, 1, 0)$ and $\widetilde{h} := (0, 3, 0)$.*

Proof. Under either action, the probabilities of entering m and z at the next step are equal. The first-entry calculation of Example 1 therefore gives $\mathbb{P}_s(\tau_m < \infty) \le 1/2$ under every controller strategy. The sample average converges to $\mathbf{1}_{\{\tau_m < \infty\}}$, and always using p^1 attains absorption probability $1/2$. This proves the true gain, including against history-dependent control. For the incorrect pair,

$$(p^0)^\top \widetilde{g} = 3/4, \qquad (p^1)^\top \widetilde{g} = 5/8, \qquad (p^0)^\top \widetilde{h} = 0, \qquad (p^1)^\top \widetilde{h} = 3/4.$$

The gain maximum is attained only by p^0, whereas the unrestricted bias maximum uses p^1. Hence the unrestricted equations hold, but the gain-active bias equation at s requires $3/4 + 0 = 0 + 0$, which is false. This shows that the controller restriction is necessary independently of nature's restriction. □

J PROOFS OF THEOREMS 5.1 AND 5.2: FINITE BELLMAN SOLVABILITY

Theorem J.1 and Corollary J.2 establish the fixed-policy criterion in Theorem 5.1. Theorems J.4 and J.5 establish the two equivalent criteria in Theorem 5.2. Appendix M gives structural sufficient conditions and examples in which solvability fails.

We distinguish two questions. For a prescribed vector g satisfying the gain equation $g = \widehat{T}g$, does the bias equation $g + h = L_g h$ have a finite solution? If the gain is also unknown, does any finite pair exist? The first question requires a fixed point of $K_g := L_g - g$. For the second, an additive eigenpair of K_g suffices, because its scalar eigenvalue can be incorporated into the gain. We first state the one-player result needed for policy evaluation and then construct the corresponding two-player criterion.

J.1 THE CLASSICAL CRITERION AND FIXED-POLICY SPECIALIZATION

Consider a finite-state minimizing MDP. At state i, its action set B_i is a nonempty compact metric space. An action $b \in B_i$ gives reward $c_i(b)$ and transition row $p_i(b)$, both continuous in b. Define the stationary deterministic policy space by $\mathcal{Q} := \prod_i B_i$. For $q \in \mathcal{Q}$, let P_q have row $p_i(q_i)^{\mathsf{T}}$ and define $c_i^q := c_i(q_i)$. Using the finite-chain quantities from Lemma C.2, set

$$\begin{aligned} \eta(q) &:= P_q^{\infty} c^q, & \gamma_i &:= \inf_{q \in \mathcal{Q}} \eta_i(q), \\ \mathcal{Q}_* &:= \{q \in \mathcal{Q} : \eta(q) = \gamma\}, & w(q) &:= Z_{P_q}\big(c^q - \eta(q)\big). \end{aligned} \tag{69}$$

Thus γ takes a separate infimum at each initial state, whereas membership in $\mathcal{Q}_*$ requires one policy to attain all of these infima. Bounded rewards ensure that γ is finite, but compactness alone does not ensure that $\mathcal{Q}_*$ is nonempty. The vector $w(q)$ uses the fixed normalization $P_q^{\infty} w(q) = 0$.

Theorem J.1. *The coupled equations*

$$z_i = \min_{b \in B_i} p_i(b)^{\mathsf{T}} z, \tag{70}$$

$$z_i + h_i = \min_{b \in B_i :\, p_i(b)^{\mathsf{T}} z = z_i} \{c_i(b) + p_i(b)^{\mathsf{T}} h\} \tag{71}$$

have a finite solution if and only if

$$\mathcal{Q}_* \neq \varnothing \quad \textit{and} \quad \exists B < \infty : \quad w(q) \geq -B\mathbf{1} \quad \textit{for every } q \in \mathcal{Q}_*. \tag{72}$$

Every solution satisfies $z = \gamma$. The bound in equation 72 is componentwise and must hold uniformly over all simultaneously gain-optimal stationary deterministic policies.

This is the discrete-time minimization form of Schweitzer (1985, Theorem 1). The proof below checks the hypotheses, the reward sign, and the canonical normalization explicitly.

Proof. **Step 1: apply the published criterion after changing the reward sign.** Replace c by $\widetilde{c} := -c$ and maximize rewards. Regard each transition as having holding time one. In the semi-Markov notation of the cited theorem, the holding-time matrix is then $H_q := P_q$ and the mean holding-time vector is $\mathbf{1}$. Its assumptions follow from the finite state space, compact metric action sets, continuous rewards and transitions, and the strictly positive constant holding time.

For each stationary policy q, the transformed gain is $\widetilde{\eta}(q) = -\eta(q)$. Since $P_q \eta(q) = \eta(q)$, its canonical bias is

$$\widetilde{w}(q) := Z_{P_q}\big(-c^q - H_q \widetilde{\eta}(q)\big) = -Z_{P_q}\big(c^q - \eta(q)\big) = -w(q).$$

The transformed componentwise optimal gain is $-\gamma$, and its simultaneously optimal policies are exactly $\mathcal{Q}_*$. The cited theorem requires this set to be nonempty and the vectors $\widetilde{w}(q)$ to have a common upper bound. These requirements are precisely equation 72.

Step 2: translate the functional equations. The maximizing equations in the cited theorem are

$$\widetilde{z}_i = \max_{b \in B_i} p_i(b)^{\mathsf{T}} \widetilde{z}, \qquad \widetilde{h}_i = \max_{b \in \widetilde{L}_i} \{-c_i(b) - p_i(b)^{\mathsf{T}} \widetilde{z} + p_i(b)^{\mathsf{T}} \widetilde{h}\},$$

where $\widetilde{L}_i$ is the set of maximizers in the first equation. Set $z := -\widetilde{z}$ and $h := -\widetilde{h}$. The first equation becomes equation 70, and $\widetilde{L}_i = \{b \in B_i : p_i(b)^\mathsf{T} z = z_i\}$. On this set, the second equation becomes

$$-h_i = \max_{b \in \widetilde{L}_i} \{-c_i(b) + z_i - p_i(b)^\mathsf{T} h\} = z_i - \min_{b \in \widetilde{L}_i} \{c_i(b) + p_i(b)^\mathsf{T} h\}.$$

This is equation 71. The published theorem therefore gives both directions of the solvability criterion.

Step 3: identify the gain of any solution. Let (z, h) solve equation 70–equation 71. At each state, the gain-minimizing set is nonempty and compact. Choose an action minimizing the bias expression on that set, and let q be the resulting policy. It satisfies

$$P_q z = z, \qquad (I - P_q)h = c^q - z.$$

The first identity gives $P_q^\infty z = z$. Multiplying the second by P_q^∞ gives $z = P_q^\infty c^q = \eta(q)$, and hence $z \geq \gamma$.

For the opposite inequality, fix any $q' \in \mathcal{Q}$ and abbreviate $P := P_{q'}$ and $\Gamma := P^\infty$. The gain equation gives $d := Pz - z \geq 0$. Consequently, $z \leq Pz \leq P^2 z \leq \cdots$, and taking Cesàro averages gives $z \leq \Gamma z$. Also, $\Gamma d = 0$. By Lemma C.2, d vanishes at recurrent states of P. At each such state, the action q'_i therefore satisfies $p_i(q'_i)^\mathsf{T} z = z_i$ and is feasible in the bias minimum. Thus the vector

$$e := c^{q'} + Ph - h - z$$

is nonnegative on all recurrent states. The projector Γ has zero columns at transient states, so $\Gamma e \geq 0$, regardless of the values of e at transient states. Since $\Gamma(P - I) = 0$, this yields $\eta(q') \geq \Gamma z \geq z$. Taking the componentwise infimum over q' gives $\gamma \geq z$. Hence $z = \gamma$. □

Remark 12. *Changing the reward sign sends (η, w) to $(-\eta, -w)$. It therefore changes the maximizing theorem's uniform upper bound into the lower bound used here. The normalization $P_q^\infty w(q) = 0$ is essential: choosing a different additive constant for each policy would change the family whose boundedness is being tested.*

Remark 13. *If a policy q attains the complete vector γ, then $P_q\gamma = P_q P_q^\infty c^q = \gamma$. Once γ is known to solve the gain equation, every action selected by q is therefore gain minimizing, including actions at transient states. Attainment from only one initial state does not imply this property.*

Corollary J.2. *Fix $\pi \in \Pi_S$. Apply Theorem J.1 with $B_i := \mathcal{U}_i^\pi$ from equation 20, $p_i(b) := b$, and $c_i(b) := r_i^\pi$. Its equations are the fixed-policy certificate equation 66–equation 67. For deterministic π, they reduce to*

$$g_i^\pi = \min_{p \in \mathcal{U}_{i,\pi(i)}} p^\mathsf{T} g^\pi, \qquad g_i^\pi + h_i^\pi = r_{i,\pi(i)} + \min_{p \in F_{i,\pi(i)}(g^\pi)} p^\mathsf{T} h^\pi.$$

These equations are solvable exactly when a stationary nature kernel attains the complete worst-gain vector and all such kernels have a common componentwise lower bound on their canonical biases. Every solution has gain g^π.

Proof. The effective row sets $\mathcal{U}_i^\pi$ are nonempty and compact, and their reward and transition data are continuous. A policy of the one-player model chooses one effective row from each $\mathcal{U}_i^\pi$. As established after equation 20, these choices induce exactly the transition matrices induced by full nature selectors: the map

$$q \longmapsto \Big(\sum_a \pi(a \mid i) q_{ia}\Big)_i$$

is onto the product of effective row sets. Because π is fixed, both descriptions also use the same reward vector r^π. They therefore give the same gain vectors and canonical biases. In particular, the vector γ in equation 69 equals γ^π in Theorem 5.1, and equation 29 identifies it with g^π.

Independent minimization over the actionwise rows gives

$$\min_{b \in \mathcal{U}_i^\pi} b^\mathsf{T} z = \sum_a \pi(a \mid i) m_{ia}(z).$$

Thus equation 70 is equation 66. Suppose this equation holds. Each term $p_a^\mathsf{T} z - m_{ia}(z)$ is nonnegative. Their weighted sum is zero exactly when every positive-weight row belongs to $F_{ia}(z)$. Consequently,

$$\min_{b \in \mathcal{U}_i^\pi : b^\mathsf{T} z = z_i} \{r_i^\pi + b^\mathsf{T} h\} = r_i^\pi + \sum_a \pi(a \mid i) \min_{p \in F_{ia}(z)} p^\mathsf{T} h.$$

Actions with zero probability do not affect either side. This proves that equation 71 is equation 67. Theorem J.1 now gives the claimed criterion and gain identification. For deterministic π, the effective row set is simply $\mathcal{U}_{i,\pi(i)}$. □

J.2 Fixed points, bounded orbits, and barriers

For any z satisfying $z = \widehat{T}z$, define $K_z(x) := L_z x - z$. This map is monotone and additively homogeneous, meaning that $x \le y$ implies $K_z x \le K_z y$ and $K_z(x + c\mathbf{1}) = K_z x + c\mathbf{1}$ for every scalar c. The next lemma supplies the fixed-point argument used below. An orbit means the sequence $x, Fx, F^2x, \ldots$, with $F^0x := x$.

Lemma J.3. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be monotone and additively homogeneous. The following statements are equivalent:*

1. *There exist $h \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ such that $Fh = h + \lambda\mathbf{1}$.*
2. *$\sup_{N\ge 0} \mathrm{sp}(F^N 0) < \infty$.*
3. *There exist $\ell, u \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ such that $\ell + \lambda\mathbf{1} \le F\ell$ and $Fu \le u + \lambda\mathbf{1}$.*

Moreover, F has a fixed point if and only if one of its orbits is bounded in the supremum norm.

The equivalence between (1) and (2) is Gaubert & Gunawardena (2004, Theorem 9), and the final assertion is their Lemma 3 with zero growth rate. We provide a self-contained proof, including the barrier formulation in (3).

Proof. **Step 1: continuity and nonexpansiveness.** For any $x, y \in \mathbb{R}^n$, define $a := \min_i(x_i - y_i)$ and $b := \max_i(x_i - y_i)$. The inequalities $y + a\mathbf{1} \le x \le y + b\mathbf{1}$ imply

$$Fy + a\mathbf{1} \le Fx \le Fy + b\mathbf{1}.$$

Therefore $\|Fx - Fy\|_\infty \le \|x - y\|_\infty$ and $\mathrm{sp}(Fx - Fy) \le \mathrm{sp}(x - y)$. In particular, F is continuous. Such a map is often called topical.

Step 2: an additive eigenpair gives bounded span. If $Fh = h + \lambda\mathbf{1}$, homogeneity gives $F^N h = h + N\lambda\mathbf{1}$. By Step 1,

$$\|F^N 0 - (h + N\lambda\mathbf{1})\|_\infty \le \|h\|_\infty.$$

Subtracting a scalar does not change the span, so $\mathrm{sp}(F^N 0) \le \mathrm{sp}(h) + 2\|h\|_\infty$ for every N. This proves (1)⇒(2).

Step 3: bounded span determines a scalar drift. Assume (2), and define

$$K := \sup_{N\ge 0} \mathrm{sp}(F^N 0), \quad x_N := F^N 0, \quad a_N := \max_i (x_N)_i, \quad b_N := \min_i (x_N)_i.$$

Since $b_M\mathbf{1} \le x_M \le a_M\mathbf{1}$, monotonicity and homogeneity give

$$x_N + b_M\mathbf{1} \le x_{N+M} \le x_N + a_M\mathbf{1}.$$

Hence $a_{N+M} \le a_N + a_M$ and $b_{N+M} \ge b_N + b_M$. Also, nonexpansiveness yields $\|x_N\|_\infty \le N\|F0\|_\infty$. For a fixed integer $m \ge 1$, write $N = km + r$ with $0 \le r < m$. Then $a_N \le k a_m + a_r$, so $\limsup_N a_N/N \le a_m/m$. Taking the infimum over m proves the first limit below, and applying the same argument to $-b_N$ proves the second:

$$\lim_{N\to\infty} \frac{a_N}{N} = \inf_{N\ge 1} \frac{a_N}{N}, \qquad \lim_{N\to\infty} \frac{b_N}{N} = \sup_{N\ge 1} \frac{b_N}{N}.$$

Because $a_N - b_N \le K$, these limits are equal. Denote their common value by λ. The infimum and supremum formulas imply $b_N \le N\lambda \le a_N$, and consequently

$$\|x_N - N\lambda\mathbf{1}\|_\infty \le K \qquad (N \ge 0).$$

For $G(x) := F(x) - \lambda\mathbf{1}$, homogeneity gives $G^N 0 = x_N - N\lambda\mathbf{1}$. Thus the orbit of 0 under G is bounded in the supremum norm.

Step 4: the bounded centered orbit gives a fixed point. Set $y_N := G^N 0$ and define coordinatewise

$$v_N := \inf_{k \geq N} y_k, \qquad v := \lim_{N \to \infty} v_N.$$

These vectors are finite because the sequence (y_N) is bounded. For every $k \geq N$, the inequality $v_N \leq y_k$ gives $Gv_N \leq y_{k+1}$. Taking the infimum over $k \geq N$ gives $Gv_N \leq v_{N+1}$. Continuity now implies $Gv \leq v$.

It follows by monotonicity that $v, Gv, G^2 v, \ldots$ is coordinatewise decreasing. This sequence is also bounded, because

$$\|G^N v - G^N 0\|_\infty \leq \|v\|_\infty$$

and $(G^N 0)$ is bounded. Hence $G^N v$ converges to a finite vector h. Passing to the limit in the iteration gives $Gh = h$, or $Fh = h + \lambda \mathbf{1}$. Steps 3–4 prove (2)⇒(1).

Step 5: equivalence with two barriers. Statement (1) gives (3) by taking $\ell = u = h$. Conversely, assume (3) and define $G(x) := F(x) - \lambda \mathbf{1}$. Then $\ell \leq G\ell$ and $Gu \leq u$. The barriers need not initially be ordered. Define

$$s := \max_i (\ell_i - u_i), \qquad u' := u + s\mathbf{1}.$$

Then $\ell \leq u'$, and homogeneity gives $Gu' \leq u'$. Starting from $h_0 := \ell$, define $h_{k+1} := Gh_k$. Induction using the two barrier inequalities gives

$$\ell = h_0 \leq h_1 \leq h_2 \leq \cdots \leq u'.$$

The sequence converges coordinatewise to a finite vector h, and continuity gives $Gh = h$. This proves (3)⇒(1).

Step 6: fixed points and supremum-norm boundedness. If $Fh = h$, nonexpansiveness gives $\|F^N x - h\|_\infty \leq \|x - h\|_\infty$ for every x and N, so every orbit is bounded. Conversely, if the orbit of some x is bounded, then the orbit of 0 is bounded because $\|F^N x - F^N 0\|_\infty \leq \|x\|_\infty$. Statement (2) therefore holds and gives an additive eigenpair (h, λ). Nonexpansiveness also makes the orbit of h bounded. Since that orbit is $h + N\lambda \mathbf{1}$, it can be bounded only when $\lambda = 0$. Thus F has a fixed point. □

Bounded span allows a scalar linear drift in the iterates. Bounded supremum norm forces that drift to vanish. This distinction is necessary when the gain is prescribed.

Theorem J.4. *For a prescribed g satisfying $g = \widehat{T} g$, the following statements are equivalent:*

1. *There exists a finite h such that $g + h = L_g h$.*
2. $\sup_{N \geq 0} \|K_g^N 0\|_\infty < \infty$.
3. *There exist finite ℓ, u such that $\ell \leq K_g \ell$ and $K_g u \leq u$.*

When the gain is unspecified, a finite gain-bias pair exists if and only if there is some z satisfying $z = \widehat{T} z$ for which $\sup_{N \geq 0} \operatorname{sp}(K_z^N 0) < \infty$. Equivalently, some such K_z has an additive eigenpair $K_z h = h + \lambda \mathbf{1}$. In that case, setting $g := z + \lambda \mathbf{1}$ gives a finite Bellman pair (g, h).

Proof. For prescribed g, the bias equation is exactly $K_g h = h$. The map K_g is monotone and additively homogeneous. The final assertion of Lemma J.3 gives (1)⇔(2). The barrier construction in Step 5 of that lemma, with $\lambda = 0$, gives (3)⇒(1). Taking $\ell = u = h$ gives (1)⇒(3).

For the unspecified-gain statement, suppose $z = \widehat{T} z$ and $K_z h = h + \lambda \mathbf{1}$. Every transition row satisfies $p^\mathsf{T} \mathbf{1} = 1$. Therefore adding $\lambda \mathbf{1}$ to z adds the same scalar to every continuation gain. It preserves both the maximizing controller actions and the minimizing nature rows:

$$\widehat{T}(z + \lambda \mathbf{1}) = z + \lambda \mathbf{1}, \qquad A_{z + \lambda \mathbf{1}}(i) = A_z(i), \qquad F_{ia}(z + \lambda \mathbf{1}) = F_{ia}(z).$$

Thus $L_{z + \lambda \mathbf{1}} = L_z$. For $g := z + \lambda \mathbf{1}$, it follows that $K_g h = K_z h - \lambda \mathbf{1} = h$. Conversely, a finite Bellman pair (g, h) gives the additive eigenpair $K_g h = h$ with eigenvalue zero. Lemma J.3 equates the existence of an additive eigenpair with bounded span of the orbit. Only scalar shifts are used in this argument. A nonconstant vector shift can change the active sets and cannot be substituted for them. □

Remark 14. *Span boundedness for a prescribed g does not establish solvability at that same gain, because the resulting additive eigenvalue may be nonzero. For example, in a one-state model with reward r, every scalar z solves the gain equation and $K_z(x) = x + r - z$. Every orbit has span zero, but K_z has a fixed point only when $z = r$.*

J.3 Stationary tangent saddle and canonical-bias envelope

Fix $g = \widehat{T}g$ and define $c_{ia} := r_{ia} - g_i$. The allowed controller actions are $A_g(i)$, and the allowed nature rows after action a are $F_{ia}(g)$. Define

$$\Pi_g := \prod_i A_g(i), \qquad \mathcal{Q}_g := \prod_i \prod_{a \in A_g(i)} F_{ia}(g), \qquad \mathcal{Q}_g^\pi := \prod_i F_{i,\pi(i)}(g).$$

A full plan $q \in \mathcal{Q}_g$ specifies a row for every active action. A restricted reply $q \in \mathcal{Q}_g^\pi$ specifies only the rows used by the fixed controller π. In either case, let $P_{\pi,q}$ be the induced transition matrix. Its row at state i is $q_{i,\pi(i)}^\mathsf{T}$ for a full plan and q_i^T for a restricted reply. Define

$$c_i^\pi := c_{i,\pi(i)}, \qquad \eta^{\pi,q} := P_{\pi,q}^\infty c^\pi, \qquad w^{\pi,q} := Z_{P_{\pi,q}} c^\pi \quad \text{when } \eta^{\pi,q} = 0.$$

Because π and q use only gain-active actions and rows, $P_{\pi,q} g = g$. The vector $\eta^{\pi,q}$ is therefore the original gain of this pair minus g. In particular, zero centered gain means that the pair attains the candidate gain g. All inequalities below are componentwise.

Theorem J.5. *The equation $K_g h = h$ has a finite solution if and only if there exist $\bar\pi \in \Pi_g$, a full plan $\bar q \in \mathcal{Q}_g$, and $B < \infty$ such that*

$$\eta^{\bar\pi,q} \geq 0 \qquad \textit{for every } q \in \mathcal{Q}_g^{\bar\pi}, \tag{73}$$

$$\eta^{\pi,\bar q} \leq 0 \qquad \textit{for every } \pi \in \Pi_g, \tag{74}$$

$$w^{\bar\pi,q} \geq -B\mathbf{1} \qquad \textit{for every } q \in \mathcal{Q}_g^{\bar\pi} \textit{ with } \eta^{\bar\pi,q} = 0. \tag{75}$$

The first condition fixes one controller and considers all nature replies to it. The second fixes one full nature plan and considers all active controllers. The third requires a uniform canonical-bias bound for all zero-gain replies to the controller selected in the first condition.

Proof. We first derive the three stationary conditions from a common bias. For the converse, we construct a lower barrier by fixing the controller and an upper barrier by fixing nature's full plan.

Step 1: a common bias gives both gain inequalities. Suppose $K_g h = h$. For each state, choose $\bar\pi(i)$ maximizing the action expression in $(K_g h)_i$. For every active action, choose $\bar q_{ia}$ minimizing the row expression. These choices exist because the action sets are finite and the row sets are nonempty and compact. At the selected action, the minimum equals h_i, so every feasible row satisfies $c_{i,\bar\pi(i)} + p^\mathsf{T} h \geq h_i$. For any active action, its minimum is at most h_i, and $\bar q_{ia}$ attains that minimum. Consequently,

$$c^{\bar\pi} + P_{\bar\pi,q} h - h \geq 0 \quad (q \in \mathcal{Q}_g^{\bar\pi}), \qquad c^\pi + P_{\pi,\bar q} h - h \leq 0 \quad (\pi \in \Pi_g).$$

Multiply each inequality by the corresponding nonnegative Cesàro projector. The projector annihilates $(P - I)h$, leaving precisely equation 73 and equation 74.

Step 2: the same bias bounds every zero-gain reply. Fix $q \in \mathcal{Q}_g^{\bar\pi}$ with $\eta^{\bar\pi,q} = 0$, and define

$$P := P_{\bar\pi,q}, \qquad \Gamma := P^\infty, \qquad d := c^{\bar\pi} + Ph - h.$$

Step 1 gives $d \geq 0$, and $\Gamma d = \Gamma c^{\bar\pi} = 0$. By Lemma C.2, $Z_P d \geq 0$. The identity $Z_P(I - P) = I - \Gamma$ then gives

$$w^{\bar\pi,q} = Z_P\big((I - P)h + d\big) = (I - \Gamma)h + Z_P d \geq h - \Gamma h.$$

Each coordinate of Γh is a convex combination of the coordinates of h, because Γ is stochastic. Therefore $h - \Gamma h \geq -\operatorname{sp}(h)\mathbf{1}$. This proves equation 75 with the same constant $B := \operatorname{sp}(h)$ for every zero-gain reply.

Step 3: the stationary conditions give a lower barrier. Assume equation 73–equation 75. Restrict $\bar q$ to the actions of $\bar\pi$. At this pair, the first two conditions give both $\eta^{\bar\pi,\bar q} \geq 0$ and $\eta^{\bar\pi,\bar q} \leq 0$, so the centered gain is zero at every state.

Now fix $\bar{\pi}$. Nature faces a minimizing MDP with row set $F_{i,\bar{\pi}(i)}(g)$ and reward $c_{i,\bar{\pi}(i)}$ at state i. All of its stationary gains are nonnegative by equation 73, while the restricted $\bar{q}$ attains zero simultaneously at all states. Its componentwise minimal gain is therefore the zero vector. Its simultaneously optimal policies are exactly the replies appearing in equation 75. That condition supplies the canonical-bias bound required by Theorem J.1.

The theorem gives a finite bias ℓ with zero gain. Since every probability row has inner product zero with the zero vector, the bias equation retains every row in the fixed-controller problem:

$$\ell_i = \min_{p \in F_{i,\bar{\pi}(i)}(g)} \{c_{i,\bar{\pi}(i)} + p^{\mathsf{T}} \ell\}.$$

The maximization in K_g includes $\bar{\pi}(i)$, so $\ell \le K_g \ell$.

Step 4: the full nature plan gives an upper barrier. Fix $\bar{q}$. The controller now faces a finite-action maximizing MDP with rewards c_{ia} and rows $\bar{q}_{ia}$. By equation 74, all deterministic stationary policies have centered gain at most zero, and Step 3 shows that $\bar{\pi}$ attains zero simultaneously at every state.

Apply Theorem J.1 after changing the rewards to $-c$. The minimizing gain vector is zero and is attained by $\bar{\pi}$. The required bias bound is automatic, because there are only finitely many deterministic stationary policies and each has a finite canonical bias. This is also the usual finite-action multichain existence result (Puterman, 2014, Chapter 9). Let v be the resulting minimizing bias and set $u := -v$. The gain is zero, so every active controller action remains in the bias equation. Reversing the reward sign gives

$$u_i = \max_{a \in A_g(i)} \{c_{ia} + \bar{q}_{ia}^{\mathsf{T}} u\}.$$

For each active action, $\bar{q}_{ia}$ is feasible in the row minimum defining K_g. It follows that

$$(K_g u)_i = \max_{a \in A_g(i)} \min_{p \in F_{ia}(g)} \{c_{ia} + p^{\mathsf{T}} u\} \le \max_{a \in A_g(i)} \{c_{ia} + \bar{q}_{ia}^{\mathsf{T}} u\} = u_i.$$

This step uses a full nature plan, since a row is needed for every controller action in the maximum.

Step 5: combine the two barriers. Steps 3–4 give $\ell \le K_g \ell$ and $K_g u \le u$. Theorem J.4 therefore gives a finite fixed point h of K_g. The original stationary pair $(\bar{\pi}, \bar{q})$ need not satisfy both Bellman equalities at this h. A pair that does satisfy them is obtained by selecting maximizing actions and minimizing rows at h, as in Step 1. □

K Geometry and structure of Bellman certificates

Section 5 characterizes when a finite Bellman bias exists. This section determines which controller-nature pairs admit a common bias, how small its span can be, and how all biases compatible with a fixed pair can be described. We answer these questions using the lower and upper Bellman inequalities together. We then give a quantitative sufficient condition for feasibility and a finite linear-program formulation when the gain faces are given by finite lists of rows.

K.1 Common biases and the mixed-flow characterization

Fix $g \in \mathbb{R}^n$ satisfying the recession equation $g = \widehat{T} g$, and define the centered rewards $c_{ia} := r_{ia} - g_i$. The state set has $n \ge 1$ elements, each controller action set is finite and nonempty, and every ambiguity set is nonempty and compact. Recall the selector sets

$$\Pi_g := \prod_i A_g(i), \qquad \mathcal{Q}_g := \prod_i \prod_{a \in A_g(i)} F_{ia}(g).$$

The gain equation ensures that the active action sets are nonempty. Each gain face $F_{ia}(g)$ is compact, and each of its rows satisfies $p^\top g = g_i$ when $a \in A_g(i)$. A pair $(\pi, q) \in \Pi_g \times \mathcal{Q}_g$ chooses one active controller action at each state and a nature row for *every* active state-action pair. Rows for unselected controller actions are needed to check whether the controller can profit by deviating. All optimizations over selector pairs in this section use this full product.

Define $\mathcal{H}_{\pi q}$ as the set of $h \in \mathbb{R}^n$ satisfying

$$\begin{aligned} h_i &\le c_{i,\pi(i)} + p^\top h && \text{for every } i \text{ and } p \in F_{i,\pi(i)}(g), \\ c_{ia} + q_{ia}^\top h &\le h_i && \text{for every } i \text{ and } a \in A_g(i). \end{aligned} \tag{76}$$

The first family says that the chosen controller action gives a Bellman value at least h_i against every row in its gain face. The second says that nature can keep the value of each active action at most h_i using its selected row. We call h a *common bias* for (π, q) when it satisfies both families. At the selected action and selected row, the two inequalities give the equality $h_i = c_{i,\pi(i)} + q_{i,\pi(i)}^\top h$. We prove below that the union of these common-bias sets is exactly $\mathrm{Fix}(K_g)$. The verification result then identifies the gain as $g^\star$ and gives the associated stationary guarantees.

Writing equation 76 as linear inequalities $f^\top h \le b$ gives the compact generator set, with e_i the ith coordinate vector,

$$\begin{aligned} G_{\pi q} := \{&(e_i - p, c_{i,\pi(i)}) : i = 1, \dots, n,\ p \in F_{i,\pi(i)}(g)\} \\ &\cup \{(q_{ia} - e_i, -c_{ia}) : i = 1, \dots, n,\ a \in A_g(i)\}, \end{aligned} \tag{77}$$

and the cone $C_{\pi q} := \mathrm{cone}(G_{\pi q})$ of finite nonnegative combinations, including zero. A point (f, b) of this cone combines the original constraints into $f^\top h \le b$. The vector f records the remaining statewise imbalance, and it satisfies $\mathbf{1}^\top f = g^\top f = 0$ because every generator flow does. The scalar b is the corresponding signed combination of centered rewards. These combinations use the inequalities of both players.

Define

$$\beta(\pi, q) := \sup_{(f,b) \in C_{\pi q}} \frac{[-b]_+}{\|f\|_1}, \tag{78}$$

where $[x]_+ := \max\{x, 0\}$. A zero denominator gives $+\infty$ when $b < 0$ and zero when $b \ge 0$. By construction,

$$\mathcal{H}_{\pi q} = \{h \in \mathbb{R}^n : f^\top h \le b \text{ for all } (f, b) \in C_{\pi q}\}. \tag{79}$$

For a fixed pair this is a semi-infinite linear feasibility problem. Related feasibility criteria based on projected inequalities appear in Basu et al. (2013, Theorem 2.14). Here the constraints come from the two players' Bellman comparisons, and the ratio in equation 78 determines the exact minimum compatible bias span. The separation principle is standard convex geometry; we give its short form below because the cone need not be closed. The Bellman-specific content is the use of *both* deviation families and the exact compatible-span formula.

Theorem K.1. *For fixed (π, q) the following are equivalent:*

$$\mathcal{H}_{\pi q} \neq \varnothing \quad \Longleftrightarrow \quad (0, -1) \notin \mathrm{cl}(C_{\pi q}) \quad \Longleftrightarrow \quad \beta(\pi, q) < \infty. \tag{80}$$

If these conditions hold, the minimum span is attained and equals

$$\min_{h \in \mathcal{H}_{\pi q}} \mathrm{sp}(h) = 2\beta(\pi, q). \tag{81}$$

With the convention that the infimum of an empty set is $+\infty$, the full Bellman problem satisfies

$$B_g := \inf\{\mathrm{sp}(h) : K_g h = h\} = 2 \inf_{\pi, q} \beta(\pi, q). \tag{82}$$

If $B_g < +\infty$, both infima in equation 82 are minima. In particular, a finite Bellman bias exists if and only if one pair of active selectors has a finite mixed-flow ratio. Infeasibility for a fixed pair has the sparse limiting witnesses of Proposition K.4.

Interpretation and relation to solvability. The point $(0, -1)$ represents the inconsistent inequality $0 \le -1$. Its presence only in the closure is equally obstructive: the constraints $f_k^\top h \le -1$ with $f_k \to 0$ cannot all hold for a finite h. Compact generators can have a nonclosed cone, so the closure retains limiting obstructions created by vanishing transition probabilities. Excluding them gives a common finite bias. The factor two in the span formula comes from constant-shift invariance: midpoint centering gives $\inf_t \|h - t\mathbf{1}\|_\infty = \mathrm{sp}(h)/2$.

For a prescribed pair, the theorem characterizes compatibility with a common bias. Taking the union over pairs characterizes the same fixed-point existence as Theorem 5.2 and, in addition, minimizes the required span. An incompatible pair can coexist with another pair that supports a Bellman solution, even when both are average optimal; Example 3 exhibits this distinction. Full solvability with g unknown requires the condition for some $g = \widehat{T} g$, and verification identifies every successful gain with $g^\star$.

K.2 Proof of the mixed-flow theorem

We first prove the two separation statements used below.

Lemma K.2. *Let $C \subset \mathbb{R}^n \times \mathbb{R}$ be a convex cone containing zero. Then $(0,-1) \notin \mathrm{cl}(C)$ if and only if there exists $h \in \mathbb{R}^n$ such that $f^\top h \le b$ for every $(f,b) \in C$.*

Proof. Suppose first that h satisfies all the inequalities. The closed half-space

$$\{(f,b) : b - f^\top h \ge 0\}$$

contains C and therefore also its closure. It excludes $(0,-1)$, proving one direction.

Conversely, suppose $(0,-1) \notin \mathrm{cl}(C)$. Strict separation of a point from a closed convex set gives a vector $(u,\lambda) \in \mathbb{R}^n \times \mathbb{R}$ and a scalar α such that

$$u^\top f + \lambda b \ge \alpha > -\lambda \qquad ((f,b) \in \mathrm{cl}(C)).$$

Because the cone contains zero, $\alpha \le 0$, so $\lambda > 0$. Moreover, $u^\top f + \lambda b$ must be nonnegative on the cone. Otherwise, multiplying a point with a negative value by arbitrarily large positive scalars would contradict the lower bound α. Thus $u^\top f + \lambda b \ge 0$ on C. Defining $h := -u/\lambda$ and dividing this inequality by $\lambda > 0$ gives $f^\top h \le b$, as required. □

Lemma K.3. *Let $C \subset \mathbb{R}^n \times \mathbb{R}$ be a convex cone containing zero, and let $B \ge 0$. There exists h with $\|h\|_\infty \le B$ and $f^\top h \le b$ on C if and only if*

$$b \ge -B\|f\|_1 \qquad \textit{for every } (f,b) \in C. \tag{83}$$

Proof. For necessity, a feasible h satisfies

$$b \ge f^\top h \ge -\|f\|_1 \|h\|_\infty \ge -B\|f\|_1 \qquad ((f,b) \in C),$$

where the middle inequality is Hölder's inequality.

For sufficiency, assume equation 83 and define

$$E_B := \{(f,b) : b \ge B\|f\|_1\}, \qquad D := C + E_B.$$

The inequality $f^\top h \le b$ holds for every $h \in [-B,B]^n$ exactly when $(f,b) \in E_B$, because $\max_{\|h\|_\infty \le B} f^\top h = B\|f\|_1$. In particular, E_B includes the coordinate constraints (e_i, B) and $(-e_i, B)$. We will find a vector satisfying all inequalities in D, which includes both the original constraints and these coordinate bounds.

The set D is a convex cone containing zero. Write any of its points as $(f,b) = (f_1,b_1) + (f_2,b_2)$, where $(f_1,b_1) \in C$ and $(f_2,b_2) \in E_B$. Then

$$b \ge -B\|f_1\|_1 + B\|f_2\|_1 \ge -B\|f_1 + f_2\|_1 = -B\|f\|_1.$$

The second inequality follows from $\|f_1\|_1 \le \|f_1 + f_2\|_1 + \|f_2\|_1$. This bound remains true on $\mathrm{cl}(D)$ by continuity, and therefore $(0,-1) \notin \mathrm{cl}(D)$. Lemma K.2 supplies h with $f^\top h \le b$ for every $(f,b) \in D$. Since $C \subseteq D$, the original inequalities hold. Applying the same conclusion to (e_i, B) and $(-e_i, B)$ gives $-B \le h_i \le B$ for every i. This also proves the claim when $B = 0$. □

Proof of Theorem K.1. Step 1 identifies the fixed points represented by the mixed inequalities. Steps 2–4 give fixed-pair feasibility and the exact span, and Step 5 attains the optimum over selectors.

Step 1: the mixed inequalities describe the Bellman fixed points. Fix active selectors (π, q) and let $h \in \mathcal{H}_{\pi q}$. The lower generator indexed by i and $p \in F_{i,\pi(i)}(g)$ states

$$(e_i - p)^\top h \le c_{i,\pi(i)} \quad \Longleftrightarrow \quad h_i \le c_{i,\pi(i)} + p^\top h.$$

It holds for every row of the chosen action, hence for their minimum. Allowing the controller to maximize over all active actions gives

$$h_i \le \min_{p \in F_{i,\pi(i)}(g)} (c_{i,\pi(i)} + p^\top h) \le (K_g h)_i.$$

For each active action, the upper generator states

$$(q_{ia} - e_i)^\top h \le -c_{ia} \quad \Longleftrightarrow \quad c_{ia} + q_{ia}^\top h \le h_i.$$

The face minimum is at most its value at the feasible row q_{ia}, so

$$(K_g h)_i = \max_{a \in A_g(i)} \min_{p \in F_{ia}(g)} (c_{ia} + p^\top h) \le \max_{a \in A_g(i)} (c_{ia} + q_{ia}^\top h) \le h_i.$$

The two displays give $K_g h = h$. Conversely, suppose $K_g h = h$. At each state choose a maximizing $\pi(i) \in A_g(i)$, and for every active action choose $q_{ia} \in F_{ia}(g)$ minimizing $p^\top h$; finiteness and compactness give these choices. The chosen action has minimum h_i, so every row of that action satisfies the lower inequality. Each action minimum is at most the maximum h_i, so each selected minimizing row satisfies the upper inequality. Hence

$$\mathrm{Fix}(K_g) = \bigcup_{\pi, q} \mathcal{H}_{\pi q}. \tag{84}$$

This identity uses both families of inequalities with the same vector h.

Step 2: separation for a fixed pair. An inequality $f^\top h \le b$ that holds for all generators also holds for every finite nonnegative combination: if $(f, b) = \sum_{j=1}^m \lambda_j (f_j, b_j)$ with $\lambda_j \ge 0$, then $f^\top h = \sum_j \lambda_j f_j^\top h \le \sum_j \lambda_j b_j = b$. Conversely, every generator belongs to the cone. Therefore $\mathcal{H}_{\pi q}$ is exactly the set of potentials of $C_{\pi q}$, which is equation 79, and Lemma K.2 gives

$$\mathcal{H}_{\pi q} \neq \varnothing \quad \Longleftrightarrow \quad (0, -1) \notin \mathrm{cl}(C_{\pi q}).$$

The closure is needed because a potential inequality, being continuous in (f, b), also holds at limits of cone points.

Step 3: every feasible span is at least 2β. Take $h \in \mathcal{H}_{\pi q}$, let $M_h := \max_i h_i$ and $m_h := \min_i h_i$, and set $h' := h - \frac{1}{2}(M_h + m_h)\mathbf{1}$. Every generator flow has zero sum, and so does every conic combination; hence $f^\top h' = f^\top h \le b$ on the cone. The largest and smallest entries of h' are $\pm(M_h - m_h)/2$, so $\|h'\|_\infty = \mathrm{sp}(h)/2$. Hölder's inequality gives $b \ge f^\top h' \ge -(\mathrm{sp}(h)/2)\|f\|_1$ for every cone point. If $f = 0$, this forces $b \ge 0$. If $f \neq 0$ and $b < 0$, dividing by $\|f\|_1 > 0$ gives $(-b)/\|f\|_1 \le \mathrm{sp}(h)/2$. Points with $b \ge 0$ contribute zero to the ratio. Taking the supremum yields

$$2\beta(\pi, q) \le \mathrm{sp}(h).$$

In particular, a feasible potential implies $\beta(\pi, q) < \infty$.

Step 4: a finite ratio gives a feasible bias of span 2β. Define $\beta := \beta(\pi, q)$ and suppose $\beta < \infty$. For a cone point with $f \neq 0$ and $b < 0$, the definition gives $b \ge -\beta\|f\|_1$. For $b \ge 0$ the same inequality holds because its right side is nonpositive. For $f = 0$, finiteness of β excludes $b < 0$ by the stated convention. Therefore

$$b \ge -\beta\|f\|_1 \qquad \text{for all } (f, b) \in C_{\pi q}.$$

Lemma K.3 supplies a feasible h with $\|h\|_\infty \le \beta$. Combining with Step 3,

$$2\beta \le \mathrm{sp}(h) \le 2\|h\|_\infty \le 2\beta.$$

Equality holds throughout. This proves feasibility, the exact minimum span, and its attainment, including the case $\beta = 0$. Together with Step 2, it also proves the three-way equivalence equation 80.

Step 5: optimization over selectors. The union identity equation 84 and the fixed-pair formula give

$$B_g = \inf_{K_g h = h} \mathrm{sp}(h) = \inf_{\pi, q} \inf_{h \in \mathcal{H}_{\pi q}} \mathrm{sp}(h) = 2 \inf_{\pi, q} \beta(\pi, q).$$

The formula includes the value $+\infty$, because an empty fixed-pair region has infinite ratio and an infinite infimum. Now suppose $B_g < \infty$. Choose fixed points h^k with $B_g \le \mathrm{sp}(h^k) \le B_g + 1/k$, and replace each by $h^k - (\min_i h_i^k)\mathbf{1}$. Additive homogeneity of K_g preserves the fixed-point equation, and now

$$0 \le h_i^k \le B_g + 1 \qquad \text{for every } i \text{ and } k.$$

A subsequence converges to some h^*. Continuity of K_g gives $K_g h^* = h^*$, and continuity of the maximum and minimum gives $\mathrm{sp}(h^*) = B_g$. Choose (π^*, q^*) from h^* as in Step 1, so that $h^* \in \mathcal{H}_{\pi^* q^*}$. Then

$$B_g \le \min_{h \in \mathcal{H}_{\pi^* q^*}} \mathrm{sp}(h) = 2\beta(\pi^*, q^*) \le \mathrm{sp}(h^*) = B_g.$$

The first inequality uses $\mathcal{H}_{\pi^* q^*} \subseteq \mathrm{Fix}(K_g)$. The last uses the feasibility of h^* for this pair. Equality throughout proves that (π^*, q^*) attains the selector infimum. In particular, this argument establishes attainment without requiring β to be continuous in the selector rows. □

K.3 SPARSE OBSTRUCTIONS

Proposition K.4. *Define $r := \dim \operatorname{span}\{f : (f, b) \in G_{\pi q}\}$. Each point of $C_{\pi q}$ is a nonnegative combination of at most $r + 1 \le n$ generators. If g is nonconstant, then $r + 1 \le n - 1$. If $\mathcal{H}_{\pi q} = \varnothing$, there is a sequence*

$$(f_k, -1) = \sum_{j=1}^{m_k} \lambda_{kj} z_{kj}, \qquad m_k \le r + 1, \quad \lambda_{kj} \ge 0, \quad z_{kj} \in G_{\pi q}, \quad \|f_k\|_1 \longrightarrow 0.$$

If a negative exactly balanced combination exists, the sequence can be taken constant. Otherwise, failure is witnessed by increasingly balanced combinations, each with at most $r + 1$ generators.

Proof. **Dimension count.** Define $L := \operatorname{span}\{f : (f, b) \in G_{\pi q}\}$, so that $\dim L = r$. Each flow f satisfies $\mathbf{1}^\top f = g^\top f = 0$. Hence $r \le n - 1$, and $r \le n - 2$ when g is nonconstant, because $\mathbf{1}$ and g are then linearly independent.

Sparsity. Every generator lies in $L \times \mathbb{R}$, a space of dimension $r + 1$. The conic Carathéodory theorem therefore gives a representation of each nonzero cone point using at most $r + 1$ generators. For completeness, start from any finite representation $\sum_j \lambda_j z_j$ with all retained coefficients positive. If more than $r + 1$ generators occur, they are linearly dependent. Choose coefficients θ_j, not all zero, with $\sum_j \theta_j z_j = 0$, and reverse their signs if necessary so that some $\theta_j > 0$. Set $t := \min_{\theta_j > 0} \lambda_j / \theta_j$. Replacing λ_j by $\lambda_j - t\theta_j$ leaves the represented point unchanged, keeps every coefficient nonnegative, and makes at least one coefficient zero. Repeating this reduction proves the bound. The argument concerns finite representations and does not require the cone to be closed. Zero has the empty representation.

Limiting witnesses. Suppose $\mathcal{H}_{\pi q} = \varnothing$. Theorem K.1 gives $(0, -1) \in \operatorname{cl}(C_{\pi q})$, so there are $(\widetilde{f}_k, \widetilde{b}_k) \in C_{\pi q}$ converging to $(0, -1)$. After discarding finitely many terms, $\widetilde{b}_k < 0$. Define $f_k := \widetilde{f}_k / (-\widetilde{b}_k)$. Positive rescaling gives

$$(f_k, -1) = \frac{(\widetilde{f}_k, \widetilde{b}_k)}{-\widetilde{b}_k} \in C_{\pi q}, \qquad \|f_k\|_1 = \frac{\|\widetilde{f}_k\|_1}{-\widetilde{b}_k} \longrightarrow 0.$$

Represent each $(f_k, -1)$ with at most $r + 1$ generators by the sparsity step. If $(0, b) \in C_{\pi q}$ for some $b < 0$, its rescaling to $(0, -1)$ can be used at every index. Otherwise no exactly balanced negative witness exists, and a limiting sequence is necessary. □

Interpretation. Sparsity bounds the number of generators in each witness. The coefficients can diverge and the rows can vary with k. The proposition therefore retains limiting obstructions without replacing the compact ambiguity by one finite list of rows.

K.4 AVERAGE OPTIMALITY AND COMMON-BIAS COMPATIBILITY

Example 3 (Separate barriers do not certify the same selectors)**.** *The states are (x, y, z, w), with one controller action per state and rewards $(0, -1, 0, 1)$. State y moves to x, and states z and w are absorbing. Let*

$$\mathcal{U}_x := \operatorname{co}\{p^0, p^1\}, \qquad p^0 := (0, 0, 1, 0)^\top, \quad p^1 := (0, 1/2, 1/2, 0)^\top.$$

For the selector $q_x = p^0$, separate lower and upper potentials exist, but no common potential exists. For $q'_x = p^1$, the minimum common bias span is two and $\beta(\pi, q') = 1$.

Proof. **Gain and centered rewards.** Every row of $\mathcal{U}_x$ has the form $(0, \lambda/2, 1 - \lambda/2, 0)^\top$ with $\lambda \in [0, 1]$. At each visit to x, the conditional probability of moving to z is at least $1/2$, even when nature chooses its row using the entire past. A move to y is followed immediately by a return to x. Thus the probability of making at least k such excursions before absorption is at most 2^{-k}. Absorption at z occurs almost surely, and the expected number of visits to y is finite. Starting at x, y, or z, the sample reward average therefore tends to zero. Starting at w, it equals one. Rewards are bounded, so the expected averages have the same limits. Hence all four payoff conventions give

$g = (0, 0, 0, 1)^\top$ under every nature strategy, and every stationary nature selector is average optimal. Every row is gain active, since $p^\top g = 0 = g_x$ for all $p \in \mathcal{U}_x$, and $c = r - g = (0, -1, 0, 0)^\top$.

Separate potentials for $q_x = p^0$. The lower potential $\ell := (-1, -2, 0, 0)^\top$ satisfies

$$c_x + (p^0)^\top \ell = 0 \geq \ell_x, \qquad c_x + (p^1)^\top \ell = -1 = \ell_x, \qquad c_y + \ell_x = -2 = \ell_y.$$

The lower inequalities at the absorbing states are equalities, and affineness in p extends the endpoint checks to every row of $\mathcal{U}_x$. Writing P_q for the selected transition matrix, the upper potential $u := (0, -1, 0, 0)^\top$ satisfies $u = c + P_q u$, so the upper inequalities hold with equality.

No common potential for $q_x = p^0$. A common potential satisfies the lower and upper inequalities for the selected row, hence equality: $h_x = h_z$ and $h_y = h_x - 1$. The lower inequality for p^1 then requires

$$h_x \leq (h_y + h_z)/2 = h_x - 1/2,$$

which is impossible. Equivalently, the mixed cone contains

$$(e_x - \tfrac{1}{2}e_y - \tfrac{1}{2}e_z, 0) + \tfrac{1}{2}(e_y - e_x, -1) + \tfrac{1}{2}(e_z - e_x, 0) = (0, -1/2).$$

The first two terms are lower generators and the last is an upper generator, so $\beta(\pi, q) = +\infty$.

The pair $q'_x = p^1$. The vector $h := \ell$ satisfies both families: the lower checks are those above, and the upper inequalities hold with equality because $c_x + (p^1)^\top \ell = -1 = \ell_x$ and $c_y + \ell_x = \ell_y$. Every common bias for q' obeys

$$h_y = h_x - 1, \qquad h_x = (h_y + h_z)/2, \qquad \text{hence } h_z = h_x + 1.$$

Its span is at least $h_z - h_y = 2$, and ℓ attains this value. Theorem K.1 therefore gives $\beta(\pi, q') = 2/2 = 1$. □

Discussion. The example has polytopic ambiguity, a nonconstant gain, and stationary gain attainment for every nature selector. Separate barriers guarantee that some Bellman fixed point exists, but they need not certify the prescribed stationary selector through a common bias. Mixing the constraints is essential both for a fixed-selector characterization and for the exact minimum-span formula.

K.5 Compatible recurrent-class offsets

For a fixed pair, the selected Poisson equation leaves one free constant per recurrent class. The mixed inequalities determine which choices of these constants are compatible with all deviations.

Write $s := (\pi, q)$, $\mathcal{H}_s := \mathcal{H}_{\pi q}$, $(P_s)_{i\cdot} := q^\top_{i,\pi(i)}$, and $c^\pi_i := r_{i,\pi(i)} - g_i$. The Cesàro projector P^∞_s and the fundamental matrix are defined in equation 21. A compatible pair must satisfy $P^\infty_s c^\pi = 0$, as shown below. For such a pair, define the canonical solution

$$h^0_s := (I - P_s + P^\infty_s)^{-1} c^\pi.$$

Let $C_{s,1}, \ldots, C_{s,m_s}$ be the recurrent classes of P_s. Let $(\Phi_s)_{ij}$ be the probability of eventually entering $C_{s,j}$ from state i, and let row j of N_s be the invariant distribution of $C_{s,j}$, extended by zero to the remaining states. Thus $\Phi_s \in \mathbb{R}^{n \times m_s}$ and $N_s \in \mathbb{R}^{m_s \times n}$. Finally, put $\mathcal{Z}_s := \{z \in \mathbb{R}^{m_s} : h^0_s + \Phi_s z \in \mathcal{H}_s\}$.

Proposition K.5. *For $g = \widehat{T} g$,*

$$\{h : K_g h = h\} = \bigcup_{s:\, P^\infty_s c^\pi = 0} \{h^0_s + \Phi_s z : z \in \mathcal{Z}_s\}. \tag{85}$$

For every pair included in the union, $\mathcal{Z}_s$ is closed and convex, possibly empty, and the map $z \mapsto h^0_s + \Phi_s z$ is one-to-one. The corresponding sets of biases in $\mathbb{R}^n$ may overlap for different pairs. For polytopic gain faces, finitely many vertex-selector pairs suffice, and their regions are polyhedral.

The vector z records one constant per recurrent class. At a transient state, the change in bias is the absorption-probability-weighted average of those constants. The mixed inequalities determine which choices remain valid against *both* players' deviations. The canonical choice $z = 0$ can therefore fail even when the same selector pair admits a compatible repair; solving for $z \in \mathcal{Z}_s$ performs

this repair, and Example 4 illustrates it. The linear-chain representation is classical (Schweitzer, 1985, Section 2, equations (2.6)–(2.9)); the additional restriction here is the common two-player certificate $\mathcal{H}_s$. Class-based descriptions of Bellman solutions have substantial precedents: Akian & Gaubert (2003, Theorem 1.1) characterize eigenspaces of convex monotone homogeneous maps using critical classes, and Zurek & Chen (2025, Lemma D.3 and Theorem D.4) study gain-direction shifts that enforce additional Bellman inequalities and can substantially increase the bias span in nominal multichain MDPs. Here the possibly nonconvex max-min operator is handled pair by pair, and the mixed inequalities impose compatibility with both players.

Proof of Proposition K.5. **Step 1: the selected Poisson equation.** The two mixed inequalities at the selected action and row force $(I - P_s)h = c^\pi$ for every $h \in \mathcal{H}_s$. Multiplying by P_s^∞ gives $P_s^\infty c^\pi = 0$, so pairs with $P_s^\infty c^\pi \neq 0$ have $\mathcal{H}_s = \varnothing$ and can be discarded. For every other pair, Lemma C.2 shows that h_s^0 solves the Poisson equation and satisfies $P_s^\infty h_s^0 = 0$.

Step 2: all solutions of the Poisson equation. The absorption representation gives $P_s^\infty = \Phi_s N_s$. Starting in a recurrent class, the chain stays in that class, so $N_s\Phi_s = I_{m_s}$; in particular, Φ_s has full column rank. A vector d is called harmonic for P_s if $P_s d = d$. Such a vector satisfies $P_s^t d = d$ for every t, so taking Cesàro averages gives $P_s^\infty d = d$. Conversely, $P_s P_s^\infty = P_s^\infty$ shows that every vector in the image of P_s^∞ is harmonic. Thus $\ker(I - P_s) = \operatorname{im} P_s^\infty$. The identity $N_s\Phi_s = I_{m_s}$ also shows that N_s maps onto $\mathbb{R}^{m_s}$. Therefore $\operatorname{im} P_s^\infty = \operatorname{im}(\Phi_s N_s) = \operatorname{im} \Phi_s$. Hence

$$(I - P_s)h = c^\pi \quad \Longleftrightarrow \quad h = h_s^0 + \Phi_s z \text{ for a unique } z \in \mathbb{R}^{m_s}.$$

The coordinates are $z = N_s h$. Indeed, $N_s h_s^0 = 0$ because $\Phi_s N_s h_s^0 = P_s^\infty h_s^0 = 0$ and Φ_s has full column rank. Each coordinate z_j is the invariant average of h over the class $C_{s,j}$.

Step 3: admissible offsets. Substituting this representation into the mixed inequalities shows which offsets are admissible:

$$\begin{aligned} (e_i - p)^\top \Phi_s z &\le c_{i,\pi(i)} - (e_i - p)^\top h_s^0 && (p \in F_{i,\pi(i)}(g)), \\ (q_{ia} - e_i)^\top \Phi_s z &\le -c_{ia} - (q_{ia} - e_i)^\top h_s^0 && (a \in A_g(i)). \end{aligned}$$

Their solution set is exactly $\mathcal{Z}_s$, an intersection of closed half-spaces, hence closed and convex. The union identity equation 84 now proves equation 85.

Step 4: polytopic gain faces. If each gain face is a polytope, the lower inequalities need only be checked at its vertices, because they are affine in p. Every minimizing upper row can also be chosen at a vertex. There are finitely many such selector pairs, and each corresponding offset region is defined by finitely many linear inequalities, hence polyhedral. □

Because every active row satisfies $p^\top \mathbf{1} = 1$ and $p^\top g = g_i$, the bias set is invariant under the addition of $u\mathbf{1} + tg$ for any $u, t \in \mathbb{R}$: indeed $K_g(h + u\mathbf{1} + tg) = K_g h + u\mathbf{1} + tg$. These two directions need not exhaust the admissible recurrent-class offsets.

Example 4 (A canonical bias can be repaired without changing the pair)**.** *There are three nominal states. At state* 1*, the controller may stay, with reward zero, or move to state* 2*, with reward* -1*. State* 2 *moves to state* 3 *with reward* 2*, and state* 3 *is absorbing with reward zero. Every stationary controller has gain* $g = 0$*. Choose the controller that stays at state* 1*. Its recurrent classes are* $\{1\}$ *and* $\{3\}$*. Writing* $z = (z_1, z_2)$ *for the offsets of these two classes,*

$$h^0 = (0, 2, 0)^\top, \qquad \Phi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \qquad h := h^0 + \Phi z = (z_1,\, 2 + z_2,\, z_2)^\top.$$

The selected Poisson equation holds for every z*. The only additional Bellman inequality comes from moving at state* 1*, namely* $-1 + h_2 \le h_1$*, or* $z_1 - z_2 \ge 1$*. Thus* $z = 0$ *fails, but the same pair admits* $z = (1, 0)^\top$ *and the exact bias* $h = (1, 2, 0)^\top$*.*

Proof. Under staying, states 1 and 3 collect zero reward forever, and state 2 collects reward 2 once before absorption. Moving at state 1 adds only the one-time reward -1. Hence all policies have zero average gain. The displayed h^0 solves the selected Poisson equation and has zero invariant average on each recurrent class, so it is the canonical solution. The Bellman equations are $h_1 = \max\{h_1, -1 + h_2\}$, $h_2 = 2 + h_3$, and $h_3 = h_3$, which give the stated offset condition. For $z = (1, 0)$ one checks $h_1 = 1 = \max\{1, -1 + 2\}$ and $h_2 = 2 = 2 + h_3$. □

K.6 A QUANTITATIVE GEOMETRIC SUFFICIENT CONDITION

Let $D_{\pi q} := \mathrm{conv}(G_{\pi q})$, and let $A_{\pi q}$ be its projection onto the flow coordinate. Both sets are compact, by compactness of $G_{\pi q}$ and the finite-dimensional convex-hull theorem. Put $L_{\pi q} := \mathrm{span}(A_{\pi q})$. At every state, the lower generator for the selected row and the upper generator for the selected action satisfy

$$\tfrac{1}{2}(e_i - q_{i\pi(i)}, c_{i\pi(i)}) + \tfrac{1}{2}(q_{i\pi(i)} - e_i, -c_{i\pi(i)}) = (0, 0).$$

Hence $0 \in D_{\pi q}$, and its flow projection contains zero.

Theorem K.6. *Fix selectors (π, q) and assume*

1. *every $(0, b) \in D_{\pi q}$ satisfies $b \geq 0$;*
2. *for some $\rho > 0$,*

$$\{f \in L_{\pi q} : \|f\|_1 \leq \rho\} \subset A_{\pi q}.$$

With $\bar{b} := \max\{[b]_+ : (f, b) \in D_{\pi q}\}$, there exists $h \in \mathcal{H}_{\pi q}$ such that

$$\mathrm{sp}(h) = 2\beta(\pi, q) \leq \frac{2\bar{b}}{\rho}. \tag{86}$$

The second assumption is equivalent to 0 belonging to the relative interior of $A_{\pi q}$. It permits multiple recurrent classes and does not require a lower bound on positive probabilities.

Proof. We show that the negative part of any centered reward in $D := D_{\pi q}$ is bounded by the size of its flow; the idea is to cancel the flow with an opposite flow and apply the first assumption to the resulting balanced mixture. Write $A := A_{\pi q}$, $L := L_{\pi q}$, and $\beta := \beta(\pi, q)$.

Step 1: the degenerate case. If $L = \{0\}$, every point of D has zero flow. The first assumption then gives $b \geq 0$ throughout D, hence throughout its cone, so $\beta = 0$, and Theorem K.1 proves the claim.

Step 2: a bound on D. Suppose $L \neq \{0\}$ and take $(f, b) \in D$. If $f = 0$, the first assumption gives $b \geq 0$. If $f \neq 0$, the vector $f' := -\rho f/\|f\|_1$ lies in L and has $\|f'\|_1 = \rho$. By the second assumption, $f' \in A$, so some b' satisfies $(f', b') \in D$. Define the positive weights

$$\alpha := \frac{\rho}{\rho + \|f\|_1}, \qquad 1 - \alpha = \frac{\|f\|_1}{\rho + \|f\|_1}.$$

They cancel the flow:

$$\alpha f + (1 - \alpha) f' = \frac{\rho f}{\rho + \|f\|_1} - \frac{\|f\|_1}{\rho + \|f\|_1} \cdot \frac{\rho f}{\|f\|_1} = 0.$$

Since D is convex, the mixture lies in D, and the first assumption gives

$$\frac{\rho b + \|f\|_1 b'}{\rho + \|f\|_1} \geq 0.$$

Multiplying by the positive denominator and dividing by $\rho > 0$ gives $b \geq -(b'/\rho)\|f\|_1$. By definition of $\bar{b}$, $b' \leq [b']_+ \leq \bar{b}$, and multiplying this upper bound by $-\|f\|_1/\rho \leq 0$ reverses it. Therefore

$$b \geq -\frac{b'}{\rho}\|f\|_1 \geq -\frac{\bar{b}}{\rho}\|f\|_1,$$

which also covers the case $f = 0$.

Step 3: transfer to the cone. Write a nonzero conic combination as

$$\sum_j \lambda_j z_j = \Lambda \sum_j \frac{\lambda_j}{\Lambda} z_j, \qquad \Lambda := \sum_j \lambda_j > 0.$$

The normalized sum lies in D. Multiplying its inequality from Step 2 by Λ preserves the direction and uses $\|\Lambda f\|_1 = \Lambda \|f\|_1$. Thus $b \geq -(\bar{b}/\rho)\|f\|_1$ holds on $C_{\pi q}$, including its zero point, and hence $\beta(\pi, q) \leq \bar{b}/\rho$. Theorem K.1 supplies a feasible bias with span exactly $2\beta(\pi, q)$, hence at most $2\bar{b}/\rho$.

Step 4: the relative-interior formulation. Since $0 \in A$, the affine hull of A equals $\mathrm{span}(A) = L$. By definition, $0 \in \mathrm{ri}(A)$ means that A contains a neighborhood of zero relative to L. In finite dimension, this is equivalent to containing a closed ℓ_1-ball in L of some positive radius; conversely, such a ball contains a relatively open neighborhood of zero. This proves the stated equivalence. □

Discussion. The two assumptions play different roles. The first rules out a negative exactly balanced mixture. The second allows a nonzero flow to be cancelled using a row mixture whose weight is proportional to the size of that flow. Together they bound how negative the reward component can be relative to the flow norm, which is exactly the bound needed for a finite $\beta(\pi, q)$. These assumptions are sufficient, not necessary. In particular, polyhedral models may be solvable even when the origin lies on the relative boundary of the flow projection.

Checking the flow-interior condition with reference flows. One way to verify the second assumption uses only finitely many reference flows. Define the stochastic matrix $Q_{i\cdot} := q_{i,\pi(i)}^\top$, and define $L_0 := \mathrm{im}((I-Q)^\top)$, the row space of $I-Q$ viewed as a subspace of column vectors. If every flow in $G_{\pi q}$ belongs to L_0, then $L_{\pi q} = L_0$ and the flow-interior condition holds. To see this, note that the flow projection contains $\pm(e_i - Q_{i\cdot}^\top)$ for every state, from the two generators of the selected row. These vectors span L_0. Choose a basis $f_1, \dots, f_r$ among them and write $\mathcal{L}\alpha := \sum_{j=1}^r \alpha_j f_j$. If $r > 0$, $\mathcal{L} : \mathbb{R}^r \to L_0$ is a linear isomorphism. Its inverse has finite norm $C_{\mathcal{L}} := \|\mathcal{L}^{-1}\|_{1\to 1} > 0$, where the domain L_0 carries the restriction of the ambient ℓ_1 norm. For $f \in L_0$ with $\|f\|_1 \le 1/C_{\mathcal{L}}$,

$$\|\mathcal{L}^{-1} f\|_1 \le C_{\mathcal{L}} \|f\|_1 \le 1 \quad \Longrightarrow \quad f \in \mathrm{conv}\{0, \pm f_1, \dots, \pm f_r\} \subseteq A_{\pi q},$$

where the implication follows by weighting the signed f_j with the absolute values of the coordinates of $\mathcal{L}^{-1} f$ and assigning any unused weight to zero, and the inclusion uses convexity of $A_{\pi q}$ and $0 \in A_{\pi q}$. If $L_0 = \{0\}$, the condition is immediate. Since all flows lie in L_0 by assumption, $L_{\pi q} = L_0$. Equivalently, each flow must annihilate every harmonic vector d with $Qd = d$, because the orthogonal complement of the row space of $I - Q$ is its null space. This condition can hold with several classwise harmonic coordinates.

K.7 Finite linear programs for polyhedral gain faces

Proposition K.7. *Suppose every gain face is the convex hull of finitely many listed rows. In computing B_g, it suffices to enumerate active controllers and full selectors that take one listed row per active state-action pair. For each enumerated pair, write its finitely many generator inequalities as $Mh \le d$. Its minimum bias span is the value of the linear program*

$$\min_{h,t} \{ t : Mh \le d, \quad 0 \le h_i \le t \ (i = 1, \dots, n), \quad t \ge 0 \}. \tag{87}$$

When this program is feasible, the same value equals

$$2 \max_{\lambda \ge 0} \{ -d^\top \lambda : \|M^\top \lambda\|_1 \le 1 \}. \tag{88}$$

If the primal is infeasible, the maximization in equation 88 is unbounded. The smallest enumerated primal value equals B_g; if all enumerated programs are infeasible, $B_g = +\infty$.

Proof. **Step 1: finite reduction.** An affine inequality holds throughout the convex hull of a finite list exactly when it holds at every listed row, and a linear minimum over that hull is attained at a listed row. Hence, in Step 1 of the proof of Theorem K.1, every Bellman fixed point admits a listed selector pair, and each such pair has finitely many inequalities $Mh \le d$.

Step 2: the primal program. Fix one pair. Every row of M is a flow with zero sum, so $M\mathbf{1} = 0$. Translating a feasible h by $-(\min_i h_i)\mathbf{1}$ therefore preserves $Mh \le d$ and gives $0 \le h_i \le \mathrm{sp}(h)$. Conversely, a feasible point of equation 87 satisfies $\mathrm{sp}(h) \le t$. This proves that equation 87 computes the minimum span.

Step 3: the dual program. By midpoint centering, half this value equals the value of the linear program

$$\min_{h,B} \{ B : Mh \le d, \ -B\mathbf{1} \le h \le B\mathbf{1}, \ B \ge 0 \}.$$

Attach nonnegative multipliers λ to $Mh \le d$ and keep the box and $B \ge 0$ as constraints. For fixed B, minimizing the Lagrangian over the box gives

$$\min_{\|h\|_\infty \le B} \{ B + \lambda^\top (Mh - d) \} = -d^\top \lambda + B\big(1 - \|M^\top \lambda\|_1\big).$$

Minimizing further over $B \ge 0$ gives $-d^\top \lambda$ when $\|M^\top \lambda\|_1 \le 1$ and $-\infty$ otherwise. Linear-programming duality therefore gives equation 88 whenever the primal is feasible. The primal

objective is bounded below by zero and above by the norm of any feasible bias. A finite linear program with a nonempty feasible region and a finite optimum attains that optimum, and linear-programming strong duality gives an attained dual optimum as well. The norm constraint is itself a finite collection of linear constraints after introducing $z \geq 0$ with $-z \leq M^\top \lambda \leq z$ and $\mathbf{1}^\top z \leq 1$.

Step 4: infeasibility and the overall minimum. If $Mh \leq d$ is infeasible, Farkas' lemma supplies $\lambda \geq 0$ with $M^\top \lambda = 0$ and $d^\top \lambda < 0$. Every positive multiple of λ is dual feasible, and its objective diverges to $+\infty$. Finally, the finite selector reduction and equation 82 identify the smallest enumerated value with B_g, including the case in which every pair is infeasible. □

Scope. This finite optimization computes the minimum bias span for a prescribed recession fixed point and listed polyhedral gain faces. Selector enumeration can be exponential. Its role is to evaluate the geometric characterization explicitly in this special case. The unknown-gain planner of Section 6 instead uses ordinary robust Bellman updates under compact ambiguity and finite Bellman solvability.

L Proof of unknown-gain anchored planning

This section proves Theorem 6.1. All vector norms are supremum norms unless stated otherwise. Fix a finite Bellman solution (g, h). Earlier results give $g = g^\star$ and show that

$$\omega_h(t) := \|T(h + tg) - h - (t+1)g\|_\infty \longrightarrow 0 \qquad (t \to \infty)$$

(Theorem 4.1 and Lemma I.1). Thus, applying T to $h + tg$ produces approximately the next point $h + (t+1)g$. We call $\omega_h(t)$ the affine defect at time t. We also use the nonexpansiveness of T, established in Lemma C.1.

The proof has three parts. We first bound an anchored iteration relative to a point that is only approximately fixed by its operator. We then apply this bound to the shifted operator $v \mapsto Tv - \widehat{g}_N$, with comparison point $h + Ng$. Finally, we use the resulting displacement and direction estimates to prove that every sufficiently late greedy controller is average optimal.

The updates specialize the approximately shifted Halpern iteration of Zurek & Chen (2025, Algorithm 1) to the robust Bellman operator. The additional issue is that a finite robust Bellman solution can have $\omega_h(t) > 0$ at every finite time. We therefore retain this defect throughout the estimates.

L.1 Halpern iteration near an approximate fixed point

Lemma L.1. *Let $F : X \to X$ be nonexpansive on a normed vector space. Fix an anchor z_0 and a comparison point z_*, and set $D := \|z_0 - z_*\|$ and $e := \|Fz_* - z_*\|$. Define*

$$z_{t+1} := \frac{2}{t+3} z_0 + \frac{t+1}{t+3} F z_t, \qquad t \geq 0.$$

Then, for all nonnegative integers t and N,

$$\|z_t - z_*\| \leq D + \frac{t}{3} e, \qquad \|Fz_N - z_N\| \leq \frac{8D}{N+3} + \frac{4}{3} e. \tag{89}$$

The comparison point z_* need not be a fixed point. Its residual e quantifies the additional error.

Proof. **Step 1: distance from the comparison point.** Subtract z_* from the update to obtain $z_{t+1} - z_* = \frac{2}{t+3}(z_0 - z_*) + \frac{t+1}{t+3}(Fz_t - z_*)$. Nonexpansiveness gives $\|Fz_t - z_*\| \leq \|Fz_t - Fz_*\| + \|Fz_* - z_*\| \leq \|z_t - z_*\| + e$:

$$\|z_{t+1} - z_*\| \leq \frac{2D}{t+3} + \frac{t+1}{t+3}\big(\|z_t - z_*\| + e\big).$$

We prove the first bound in equation 89 by induction. At $t = 0$ it is an equality. If it holds at time t, the recursion gives

$$\|z_{t+1} - z_*\| \leq \frac{2D}{t+3} + \frac{t+1}{t+3} D + \frac{t+1}{t+3}\Big(\frac{t}{3} + 1\Big) e = D + \frac{t+1}{3} e,$$

because $\frac{t+1}{t+3}\big(\frac{t}{3} + 1\big) = \frac{t+1}{3}$.

Step 2: a uniform bound on $\|Fz_t - z_0\|$. Fix a nonnegative integer N and define $M := 2D + (N/3 + 1)e$. For $0 \le t \le N$, Step 1 gives

$$\|Fz_t - z_0\| \le \|Fz_t - Fz_*\| + \|Fz_* - z_*\| + \|z_* - z_0\| \le \Big(D + \frac{t}{3}e\Big) + e + D \le M.$$

Step 3: successive differences. Define $d_t := \|z_t - z_{t-1}\|$ for $t \ge 1$. The first update gives $z_1 - z_0 = \frac{1}{3}(Fz_0 - z_0)$, so $d_1 \le M/3$. For $1 \le t \le N$, subtracting consecutive updates gives

$$z_{t+1} - z_t = \frac{t+1}{t+3}\big(Fz_t - Fz_{t-1}\big) + \frac{2}{(t+2)(t+3)}\big(Fz_{t-1} - z_0\big),$$

because $\frac{t+1}{t+3} - \frac{t}{t+2} = \frac{2}{(t+2)(t+3)} = \frac{2}{t+2} - \frac{2}{t+3}$. The first term measures the change in the operator output. The second term accounts for the change in the anchor weight. Nonexpansiveness and Step 2 therefore give

$$d_{t+1} \le \frac{t+1}{t+3} d_t + \frac{2M}{(t+2)(t+3)}.$$

We claim $d_t \le 2M/(t+2)$ for $1 \le t \le N+1$. The case $t = 1$ follows from $d_1 \le M/3 \le 2M/3$. If the claim holds at t, then

$$d_{t+1} \le \frac{t+1}{t+3} \cdot \frac{2M}{t+2} + \frac{2M}{(t+2)(t+3)} = \frac{2M}{t+3}.$$

Step 4: the residual. The update at time N gives $Fz_N - z_{N+1} = \frac{2}{N+3}(Fz_N - z_0)$, so Step 2 gives $\|Fz_N - z_{N+1}\| \le 2M/(N+3)$. Hence

$$\|Fz_N - z_N\| \le \|Fz_N - z_{N+1}\| + d_{N+1} \le \frac{4M}{N+3} = \frac{8D}{N+3} + \frac{4}{3}e.$$

Although the proof uses z_{N+1}, the residual at z_N requires only one evaluation of Fz_N. □

L.2 Gain, displacement, and direction estimates

Fix a finite Bellman solution (g, h) and define

$$A_N := \|h\|_\infty + \sum_{j=0}^{N-1} \omega_h(j), \qquad \varepsilon_N := \frac{A_N + \|h\|_\infty}{N}, \qquad N \ge 1. \tag{90}$$

Since $\omega_h(j) \to 0$, the average $N^{-1}\sum_{j=0}^{N-1} \omega_h(j)$ also tends to zero. Consequently, $A_N/N \to 0$ and $\varepsilon_N \to 0$. These quantities depend on the unknown Bellman solution and are used only in the proof.

Proposition L.2. *For every integer $N \ge 1$, the outputs of Algorithm 1 satisfy*

$$\|\widehat{g}_N - g\|_\infty \le \varepsilon_N, \tag{91}$$

$$\eta_N := \|TZ_N - Z_N - g\|_\infty \le \frac{8A_N}{N+3} + \frac{4}{3}\omega_h(N) + \frac{7}{3}\varepsilon_N, \tag{92}$$

$$\|Z_N - h - Ng\|_\infty \le A_N + \frac{N}{3}\big(\omega_h(N) + \varepsilon_N\big). \tag{93}$$

In particular, all three limits in equation 14 hold.

Proof. **Step 1: gain estimation.** The first phase starts from $x_0 = 0$ and applies $x_{j+1} = Tx_j$, so $x_j = T^j 0$. Compare these iterates with the points $h + jg$. Nonexpansiveness and the definition of the affine defect give

$$\begin{aligned}\|x_{j+1} - h - (j+1)g\|_\infty &\le \|Tx_j - T(h + jg)\|_\infty + \|T(h + jg) - h - (j+1)g\|_\infty \\ &\le \|x_j - h - jg\|_\infty + \omega_h(j).\end{aligned}$$

Starting from $\|x_0 - h\|_\infty = \|h\|_\infty$ and summing over $j = 0, \dots, N-1$ yields $\|x_N - h - Ng\|_\infty \le A_N$. Consequently,

$$\Big\|\frac{x_N}{N} - g\Big\|_\infty \le \frac{\|x_N - h - Ng\|_\infty + \|h\|_\infty}{N} \le \frac{A_N + \|h\|_\infty}{N},$$

which is equation 91.

Step 2: the second phase as Halpern iteration. The second phase is the iteration of Lemma L.1 for the map $S_N(v) := Tv - \widehat{g}_N$, with anchor $z_0 = x_N$. Subtracting a fixed vector preserves nonexpansiveness, so the lemma applies. Choose the comparison point $z_* := h + Ng$. Step 1 gives $\|z_0 - z_*\|_\infty \le A_N$, and

$$\|S_N z_* - z_*\|_\infty \le \|T(h+Ng) - h - (N+1)g\|_\infty + \|g - \widehat{g}_N\|_\infty \le \omega_h(N) + \varepsilon_N.$$

The distance from the anchor to this point is at most A_N, while its residual under S_N is at most $\omega_h(N) + \varepsilon_N$. These are exactly the two quantities required by Lemma L.1.

Step 3: displacement and direction. Lemma L.1 at time N gives

$$\|TZ_N - Z_N - \widehat{g}_N\|_\infty \le \frac{8A_N}{N+3} + \frac{4}{3}\big(\omega_h(N) + \varepsilon_N\big).$$

Adding the gain-estimation error $\|\widehat{g}_N - g\|_\infty \le \varepsilon_N$ proves equation 92. The distance bound of the same lemma proves equation 93. To see the direction limit explicitly, the triangle inequality gives

$$\left\| \frac{Z_N}{N} - g \right\|_\infty \le \frac{A_N + \|h\|_\infty}{N} + \frac{\omega_h(N) + \varepsilon_N}{3} = \frac{4}{3}\varepsilon_N + \frac{1}{3}\omega_h(N) \longrightarrow 0.$$

Every term on the right of equation 92 also tends to zero. Therefore $TZ_N - Z_N \to g$, and equation 91 gives $\widehat{g}_N \to g$. □

L.3 Gain-active action identification and average optimality

Completion of the proof of Theorem 6.1. The preceding proposition proves the three vector limits. To obtain the policy guarantee, we first show that greedy selection eventually excludes every action with $m_{ia}(g) < g_i$. For the remaining actions, the expected gain cannot decrease under any feasible nature row. This fact allows us to sum the displacement inequality along a controlled trajectory.

Step 1: gain-active identification. Define $a_N := \|Z_N/N - g\|_\infty$. Since m_{ia} is 1-Lipschitz, uniformly in (i, a),

$$\left| \frac{r_{ia} + \min_{p \in \mathcal{U}_{ia}} p^\top Z_N}{N} - m_{ia}(g) \right| \le \frac{R}{N} + a_N.$$

Let $a^* \in A_g(i)$ attain $\max_a m_{ia}(g) = g_i$. The scaled greedy score differs from $m_{ia}(g)$ by at most $R/N + a_N$ for both a^* and the selected action $\pi_N(i)$. Since the selected action has at least the score of a^*, these two errors give

$$m_{i,\pi_N(i)}(g) \ge g_i - 2(R/N + a_N). \tag{94}$$

If an inactive action exists, define

$$\Delta_A := \min_{i,a:\, m_{ia}(g) < g_i} \{g_i - m_{ia}(g)\} > 0,$$

which is positive because there are finitely many states and controller actions. Since $a_N \to 0$, for all sufficiently large N we have $2(R/N + a_N) < \Delta_A$, and every greedy action belongs to $A_g(i)$. If no inactive action exists, every action is gain active at every budget. The positive gap uses only the finite controller action sets. Nature's row sets can be infinite.

Step 2: a uniform performance bound after identification. Fix such a budget N. For each selected action and every feasible row, gain activity gives $p^\top g \ge m_{i,\pi_N(i)}(g) = g_i$. Greediness and the definition of η_N give

$$r_{i,\pi_N(i)} + p^\top Z_N - Z_{N,i} \;\ge\; (TZ_N)_i - Z_{N,i} \;\ge\; g_i - \eta_N \qquad (p \in \mathcal{U}_{i,\pi_N(i)}). \tag{95}$$

Let nature use any randomized history-dependent strategy τ. At every stage its realized row satisfies $p_t^\top g \ge g(S_t)$. The transition rule equation 17 therefore implies

$$\mathbb{E}_i^{\pi_N,\tau}[g(S_{t+1}) \mid \mathcal{G}_t] = p_t^\top g \ge g(S_t).$$

Taking expectations and applying this inequality successively gives $\mathbb{E}_i^{\pi_N,\tau} g(S_t) \ge g_i$ for every t. The same transition rule applied to Z_N turns equation 95 into

$$\mathbb{E}_i^{\pi_N,\tau} r_{S_t,\pi_N(S_t)} \ge \mathbb{E}_i^{\pi_N,\tau} g(S_t) - \eta_N + \mathbb{E}_i^{\pi_N,\tau} Z_N(S_t) - \mathbb{E}_i^{\pi_N,\tau} Z_N(S_{t+1}).$$

Summing this inequality over $t = 0, \dots, H-1$ cancels every intermediate potential term and yields

$$\mathbb{E}_i^{\pi_N,\tau} \sum_{t=0}^{H-1} r_{S_t,\pi_N(S_t)} \geq \sum_{t=0}^{H-1} \mathbb{E}_i^{\pi_N,\tau} g(S_t) - H\eta_N + Z_{N,i} - \mathbb{E}_i^{\pi_N,\tau} Z_N(S_H)$$
$$\geq H(g_i - \eta_N) - \mathrm{sp}(Z_N). \tag{96}$$

The budget N is fixed while $H \to \infty$, so the potential term divided by H vanishes even though $\mathrm{sp}(Z_N)$ may grow with N. Dividing by H and letting $H \to \infty$ gives

$$\liminf_{H\to\infty} J_H(i; \pi_N, \tau) \geq g_i - \eta_N \qquad \text{for every } \tau \in \mathcal{Q}_H.$$

Theorem 3.1 identifies $g_i^{\pi_N}$ with the infimum of this payoff over τ, and $g = g^\star \geq g^{\pi_N}$. Hence

$$0 \leq g - g^{\pi_N} \leq \eta_N \mathbf{1}. \tag{97}$$

The payoff equivalences of Theorem 3.1 transfer this guarantee to all payoff conventions used in the paper.

Step 3: eventual exact optimality. There are finitely many deterministic stationary controllers. If some are suboptimal, their errors have a positive minimum

$$\Delta_\Pi := \min_{\pi \in \Pi_D :\, g^\pi \neq g} \|g - g^\pi\|_\infty > 0.$$

For all sufficiently large N, we have $\eta_N < \Delta_\Pi$. Inequality equation 97 then rules out every suboptimal deterministic controller. Combining this threshold with the gain-active threshold from Step 1 gives one finite N_0 that works for every choice among greedy ties. If all deterministic controllers are already optimal, no policy-gap argument is needed. Thus every output π_N with $N \geq N_0$ attains the robust value g from every initial state, including against history-dependent nature. □

Rates when a defect modulus is available. Suppose a finite Bellman solution satisfies $\omega_h(t) \leq C(1+t)^{-\alpha}$ for some $C < \infty$ and $\alpha > 0$. Summing this bound in equation 90 gives $\sum_{j<N} \omega_h(j) = O(N^{1-\alpha})$, $O(\log(N+1))$, or $O(1)$ for $0 < \alpha < 1$, $\alpha = 1$, or $\alpha > 1$, respectively. Substituting into equation 91–equation 92 yields

$$\|\widehat{g}_N - g\|_\infty + \eta_N = \begin{cases} O(N^{-\alpha}), & 0 < \alpha < 1, \\ O(\log(N+1)/N), & \alpha = 1, \\ O(N^{-1}), & \alpha > 1. \end{cases} \tag{98}$$

After gain-active identification, equation 97 gives the corresponding bound on the controller loss. If the affine defect is eventually zero, its sum is finite and the same argument gives $O(1/N)$. The constants depend on the chosen solution and its defect modulus. Under the standing compactness assumptions and finite Bellman solvability, the proof guarantees only that $\omega_h(t) \to 0$. In particular, N_0 is an existential threshold. These estimates do not provide a stopping rule based only on the observable residual $\|TZ_N - Z_N - \widehat{g}_N\|_\infty$.

L.4 Finite solvability does not imply an inverse-budget rate

Proposition L.3. *For every integer $k \geq 3$, there is a three-state robust MDP with one controller action, compact convex semialgebraic row uncertainty, and a finite vector Bellman solution, such that Algorithm 1 satisfies*

$$\|\widehat{g}_N - g\|_\infty = \Theta\big(N^{-1/(k-1)}\big).$$

Consequently, finite Bellman solvability does not imply an $O(N^{-1})$ gain-estimation rate, or any fixed positive algebraic exponent throughout this class.

Proof. **Step 1: model and Bellman certificate.** Use states (x, y, z), one action per state, and rewards $(0, -1, 1)$. State y moves deterministically to x, and z is absorbing. We suppress the unique action index in this example. At x, set

$$p_k(u) := (1 - u - u^k, u, u^k), \qquad \mathcal{U}_x := \mathrm{co}\{p_k(u) : 0 \leq u \leq 1/2\}.$$

For $0 \le u \le 1/2$ the entries are nonnegative and sum to one. The generating curve is compact, so its convex hull is compact as well. The hull is also semialgebraic, meaning that it can be described by finitely many polynomial equalities and inequalities combined using unions, intersections, and complements. Indeed, Carathéodory's theorem represents every row using at most three curve points because the rows lie in a two-dimensional affine plane. The constraints

$$p = \sum_{j=1}^{3} \lambda_j p_k(u_j), \qquad \lambda_j \ge 0, \qquad \sum_{j=1}^{3} \lambda_j = 1, \qquad 0 \le u_j \le \tfrac{1}{2}$$

are polynomial constraints in p, the weights, and the curve parameters. Eliminating the auxiliary variables preserves semialgebraicity by the Tarski–Seidenberg projection theorem.

Define $g := (0, 0, 1)$ and $h := (0, -1, 0)$. At x, $p_k(u)^\top g = u^k$, and a convex combination of curve points has gain $\sum_j \lambda_j u_j^k$, which vanishes only if $u_j = 0$ whenever $\lambda_j > 0$. Hence the unique gain-minimizing row is $e_x = p_k(0)$, and the gain equation at x gives $m_x(g) = 0 = g_x$. At y and z, the gain equations are respectively $g_x = g_y$ and $g_z = g_z$. The gain-face bias equations at (x, y, z) are $0 + 0 = 0 + 0$, $0 - 1 = -1 + 0$, and $1 + 0 = 1 + 0$. Thus (g, h) solves both parts of the vector Bellman system. Theorem 4.1 identifies g with the robust gain.

Step 2: upper bound from the affine defect. The defect vanishes at y and z. At x, $[T(h + tg)]_x = \min_{p \in \mathcal{U}_x} p^\top (h + tg)$, and a linear objective has the same minimum over the convex hull as over the generating curve. Hence

$$\omega_h(t) = -\min_{0 \le u \le 1/2} \{-u + tu^k\},$$

which is nonnegative because $u = 0$ is feasible. The function $u \mapsto -u + tu^k$ is convex on $[0, \infty)$. For $t \ge 2^{k-1}/k$, its critical point $u_t := (kt)^{-1/(k-1)}$ lies in $(0, 1/2]$ and is therefore the minimizer on $[0, 1/2]$. At this point $-1 + kt u_t^{k-1} = 0$, so $tu_t^k = u_t/k$. Thus

$$\omega_h(t) = u_t - \frac{u_t}{k} = \frac{k-1}{k}(kt)^{-1/(k-1)}.$$

The finitely many integers $t < 2^{k-1}/k$ contribute a bounded amount to the accumulated defect. Since $1/(k-1) < 1$, $\sum_{j<N} \omega_h(j) = O(N^{1-1/(k-1)})$, and equation 91 gives the upper bound $O(N^{-1/(k-1)})$.

Step 3: matching lower bound from a feasible nature strategy. Fix $N \ge 2$, and let nature use the stationary row $p_k(u)$ with $u := \frac{1}{4} N^{-1/(k-1)}$. This is a feasible row because $0 < u \le 1/4$. Starting at x, define a_t, b_t, and c_t to be the probabilities of being at x, y, and z at time t, respectively. The transition rules give

$$b_t = u a_{t-1} \quad (t \ge 1), \qquad c_t = u^k \sum_{s=0}^{t-1} a_s \le t u^k.$$

For $0 \le t \le N$, these identities, together with $b_0 = c_0 = 0$, imply $b_t \le u$ and $c_t \le N u^k$. Since $u \le 1/4$ and $N u^{k-1} = 4^{-(k-1)} \le 1/16$, it follows that $a_t = 1 - b_t - c_t \ge 1 - u(1 + N u^{k-1}) \ge 1/2$. Thus $b_t = u a_{t-1} \ge u/2$ for $1 \le t \le N$, whereas $c_t \le t u^k$. Using reward -1 at y and 1 at z gives

$$\begin{aligned} \mathbb{E}_x \sum_{t=0}^{N-1} r(S_t) &= -\sum_{t=1}^{N-1} b_t + \sum_{t=1}^{N-1} c_t \\ &\le -\frac{u(N-1)}{2} + \frac{u^k N(N-1)}{2} \\ &= -\frac{u(N-1)}{2}\big(1 - 4^{-(k-1)}\big). \end{aligned}$$

The finite-horizon robust value minimizes over all admissible nature strategies, so it is at most the payoff of this particular stationary strategy. Using $g_x = 0$, $\widehat{g}_N = T^N 0/N$, and $(N-1)/N \ge 1/2$,

$$\|\widehat{g}_N - g\|_\infty \ge -\frac{(T^N 0)_x}{N} \ge \frac{1 - 4^{-(k-1)}}{16}\, N^{-1/(k-1)}.$$

This proves the matching order. For any prescribed $\alpha > 0$, choosing k with $1/(k-1) < \alpha$ rules out an $O(N^{-\alpha})$ bound for this instance, even with an instance-dependent constant. □

This example establishes a limitation of the finite-horizon gain estimator in Algorithm 1. It does not give a lower bound for every planning algorithm. A small $u > 0$ allows repeated visits to the negative-reward state y while making transition to the positive absorbing state z much rarer, of order u^k. The gain-minimizing row at $u = 0$ stays at x, but rows with small positive u create long negative transients. The resulting nonzero affine defect explains the slow gain-estimation rate.

M SUPPORTING RESULTS AND SOLVABILITY OBSTRUCTIONS

This section gives conditions under which a finite gain-bias pair exists and explains how solvability can fail. We first obtain bounded centered discounted values from assumptions on the transition rows. We then apply those assumptions to divergence balls and construct examples exhibiting the two obstructions to solvability. The final subsection records invariance, stability, and gain-only equations.

M.1 STRUCTURAL MULTICHAIN EXISTENCE REGIMES

The conditions below concern the original ambiguity sets and allow several recurrent classes. Throughout, V_ϵ denotes the unnormalized discounted value, so $V_\epsilon = T((1-\epsilon)V_\epsilon)$ for $0 < \epsilon < 1$. The first lemma explains why a bounded remainder after subtracting g/ϵ produces a Bellman bias.

Lemma M.1. *Suppose $\epsilon_k \downarrow 0$ and, for some vector g, the centered values $h_k := V_{\epsilon_k} - g/\epsilon_k$ are bounded. Then every subsequential limit h of (h_k) gives a solution $g = \widehat{T}g$, $g + h = L_g h$.*

Proof. Choose a subsequence with $h_k \to h$ and keep the same indices. Define $v_k := \epsilon_k V_{\epsilon_k} = g + \epsilon_k h_k$. Boundedness of h_k implies $v_k \to g$.

Step 1: the gain equation. The normalized discounted equation reads $(v_k)_i = \max_a\{\epsilon_k r_{ia} + (1-\epsilon_k)m_{ia}(v_k)\}$. Comparing each term with $m_{ia}(g)$ and using that m_{ia} is 1-Lipschitz,

$$\|v_k - \widehat{T}g\|_\infty \le \epsilon_k R + (1-\epsilon_k)\|v_k - g\|_\infty + \epsilon_k\|g\|_\infty \longrightarrow 0.$$

Since also $v_k \to g$, this proves $g = \widehat{T}g$.

Step 2: the bias equation. Define $t_k := \epsilon_k^{-1} - 1$ and $z_k := (1-\epsilon_k)h_k$. Then $(1-\epsilon_k)V_{\epsilon_k} = t_k g + z_k$ and $V_{\epsilon_k} = (t_k+1)g + h_k$. Subtracting $t_k g$ from the discounted equation gives

$$g + h_k = T(t_k g + z_k) - t_k g.$$

Here $t_k \to \infty$ and $z_k \to h$. Nonexpansiveness and Lemma I.1 give

$$\|T(t_k g + z_k) - t_k g - L_g h\|_\infty \le \|z_k - h\|_\infty + \|T(t_k g + h) - t_k g - L_g h\|_\infty \longrightarrow 0.$$

Taking limits gives $g + h = L_g h$. Thus a bounded sequence of centered discounted values suffices. Convergence of the entire family is unnecessary. □

Theorem M.2. *If every $\mathcal{U}_{ia}$ is a polytope, a finite vector gain-bias pair exists. Moreover, there are g, h, and t_0 such that*

$$T(tg + h) = (t+1)g + h \qquad \textit{for all } t \ge t_0. \tag{99}$$

Proof. **Step 1: an invariant half-line.** Let E_{ia} denote the finite vertex set of $\mathcal{U}_{ia}$. A linear function attains its minimum over a polytope at a vertex. Therefore $(Tx)_i = \max_a \min_{p\in E_{ia}}(r_{ia} + p^\top x)$. This finite maximum of finite minima of affine functions is piecewise affine with finitely many pieces. The operator is also nonexpansive in the supremum norm. These are the hypotheses of Kohlberg's invariant-half-line theorem (Kohlberg, 1980). That theorem gives vectors g, h and a threshold t_0 satisfying equation 99: sufficiently far along the line $tg + h$, one application of T advances the parameter by one. No condition on recurrent classes is needed.

Step 2: the half-line solves the Bellman system. Dividing equation 99 by t gives $T(tg + h)/t \to g$. The estimate $\|T(tg+h)/t - \widehat{T}g\|_\infty \le (R + \|h\|_\infty)/t$ from Lemma I.1 then gives $\widehat{T}g = g$. Subtracting tg from equation 99 and taking the tangent limit from the same lemma gives $L_g h = g + h$. Theorem I.3 consequently identifies $g = g^\star$.

Step 3: bounded discounted centering. For $\epsilon > 0$ small enough that $t := \epsilon^{-1} - 1 \geq t_0$, define $W_\epsilon := g/\epsilon + h = (t+1)g + h$. Then $(1-\epsilon)W_\epsilon = tg + (1-\epsilon)h$, and nonexpansiveness with equation 99 gives

$$\|T((1-\epsilon)W_\epsilon) - W_\epsilon\|_\infty = \|T(tg + (1-\epsilon)h) - T(tg+h)\|_\infty \leq \epsilon\|h\|_\infty.$$

To convert this residual into an error bound, compare W_ϵ with the discounted fixed point:

$$\|V_\epsilon - W_\epsilon\|_\infty \leq (1-\epsilon)\|V_\epsilon - W_\epsilon\|_\infty + \epsilon\|h\|_\infty.$$

Moving the first term on the right to the left gives $\|V_\epsilon - W_\epsilon\|_\infty \leq \|h\|_\infty$. Hence $\|V_\epsilon - g/\epsilon\|_\infty \leq 2\|h\|_\infty$ for all sufficiently small ϵ. For ϵ bounded away from zero, the bound $\|V_\epsilon\|_\infty \leq R/\epsilon$ gives bounded centering as well. This proves the boundedness used in the main text. □

The polytope reduction checks the hypotheses of Kohlberg's theorem, and the limit calculation identifies its invariant half-line with the present gain-bias equations. The finite perfect-information case is also developed in Akian et al. (2012).

Lemma M.3. *Let $\mathcal{P}$ be a compact family of finite stochastic matrices such that, for some $\delta > 0$, every entry of every $P \in \mathcal{P}$ is either zero or at least δ. Then $P \mapsto P^\infty$ is continuous on $\mathcal{P}$, and*

$$\sup_{P\in\mathcal{P},\, 0\leq\beta<1} \left\| (I-\beta P)^{-1} - \frac{P^\infty}{1-\beta} \right\|_\infty < \infty. \tag{100}$$

Proof. **Step 1: continuity of the Cesàro projector.** Define $\Gamma(P) := P^\infty$, the Cesàro limit of the powers of P. Let $P_k \to P$ in $\mathcal{P}$. An entry with $P_{ij} = 0$ satisfies $(P_k)_{ij} < \delta$ eventually, hence $(P_k)_{ij} = 0$ eventually; an entry with $P_{ij} > 0$ satisfies $(P_k)_{ij} > 0$ eventually. Thus the support gap makes every convergent sequence eventually have the support of its limit. On a fixed support, the transient set and the recurrent classes are fixed. If Q is the transient block, B_j the block of transitions from transient states into the recurrent class C_j, and ν_j the invariant distribution of C_j, the standard finite-chain decomposition gives

$$\Gamma(P)_{D,C_j} = (I-Q)^{-1}B_j\mathbf{1}\,\nu_j^\top, \qquad \Gamma(P)_{C_j,C_j} = \mathbf{1}\nu_j^\top,$$

Here D is the transient set, and all other blocks are zero. The vector $(I-Q)^{-1}B_j\mathbf{1}$ contains the probabilities of eventual absorption in C_j. Since D is transient, the spectral radius of Q is less than one, so $I-Q$ is invertible and its inverse varies continuously with P. The vector ν_j is also continuous: it is the unique solution of the stationary equations on the irreducible class C_j, normalized to have sum one. These formulas show that $\Gamma(P_k) \to \Gamma(P)$ once the supports agree. Thus Γ is continuous on $\mathcal{P}$. This is the support-gap implication in Schweitzer (1985, Theorem 2).

Step 2: a regular formula for the discounted remainder. Fix $P \in \mathcal{P}$ and write $\Gamma := \Gamma(P)$. Since $\Gamma^2 = \Gamma$, each vector has the unique decomposition $x = \Gamma x + (I-\Gamma)x$. The first term belongs to the image of Γ, and the second belongs to its kernel. The identities $P\Gamma = \Gamma P = \Gamma$ show that P preserves these two spaces. On the image of Γ, both P and Γ act as the identity. On its kernel, Γ vanishes.

Define $M_\beta := I - \beta(P-\Gamma)$ for $\beta \in [0,1]$. This matrix is the identity on the image of Γ and agrees with $I-\beta P$ on its kernel. For $\beta < 1$, $I-\beta P$ is invertible because its inverse is the convergent series $\sum_{k\geq 0}\beta^k P^k$. For $\beta = 1$, Lemma C.2 gives the invertibility of $M_1 = I - P + \Gamma$. Thus M_β is invertible throughout $[0,1]$. For $\beta < 1$, the identity $(I-\beta P)\Gamma = (1-\beta)\Gamma$ gives $(I-\beta P)^{-1}\Gamma = \Gamma/(1-\beta)$. On the kernel of Γ, the inverses of $I-\beta P$ and M_β agree. Applying these facts to the decomposition of x gives

$$(I-\beta P)^{-1} - \frac{\Gamma}{1-\beta} = (I-\beta P)^{-1}(I-\Gamma) = M_\beta^{-1}(I-\Gamma). \tag{101}$$

Step 3: the uniform bound. By Step 1, the right side of equation 101 is a continuous function of (P,β) on the compact set $\mathcal{P}\times[0,1]$, since matrix inversion is continuous on invertible matrices. It is therefore uniformly bounded, which proves equation 100. Periodicity does not affect the argument: only the eigenvalue one is removed. □

Theorem M.4. *Suppose there is $\delta > 0$ such that*

$$p_j = 0 \quad \textit{or} \quad p_j \geq \delta \qquad \textit{for every } i, a,\ p \in \mathcal{U}_{ia},\ j. \tag{102}$$

Then a finite gain-bias pair exists. More strongly, for some vector g and some finite constant C,

$$\left\| V_\epsilon - \frac{g}{\epsilon} \right\|_\infty \le C \qquad (0 < \epsilon < 1). \tag{103}$$

The same assertion holds for every fixed stationary policy. For a fixed randomized policy, the constant may depend on its positive action probabilities.

Proof. Steps 1–3 prove the bound for optimal control, and Step 4 treats a fixed randomized policy.

Step 1: a uniform bound for stationary chains. For every deterministic controller π, the induced kernel family $\{P^{\pi,q} : q \in \mathcal{Q}_S\}$ is compact and inherits the support gap, since its rows are rows of the sets $\mathcal{U}_{i,\pi(i)}$. Lemma M.3 with $\beta = 1 - \epsilon$, the bound $\|r^\pi\|_\infty \le R$, and the finiteness of Π_D give a constant $C < \infty$ such that

$$\left\| (I - (1-\epsilon)P^{\pi,q})^{-1} r^\pi - \frac{\eta^{\pi,q}}{\epsilon} \right\|_\infty \le C, \qquad \eta^{\pi,q} := (P^{\pi,q})^\infty r^\pi, \tag{104}$$

for every $\pi \in \Pi_D$, every $q \in \mathcal{Q}_S$, and every $0 < \epsilon < 1$.

Step 2: fixed-policy values. Fix $\pi \in \Pi_D$, and define $\gamma_i^\pi := \inf_{q \in \mathcal{Q}_S} \eta_i^{\pi,q}$. For every q, equation equation 104 gives

$$\frac{\eta_i^{\pi,q}}{\epsilon} - C \le [(I - (1-\epsilon)P^{\pi,q})^{-1} r^\pi]_i \le \frac{\eta_i^{\pi,q}}{\epsilon} + C.$$

Take the infimum over q in each expression. By Lemma C.1(iii), the middle infimum is $V_{\epsilon,i}^\pi$. The outer infima are $\gamma_i^\pi/\epsilon - C$ and $\gamma_i^\pi/\epsilon + C$, respectively. Hence

$$\left\| V_\epsilon^\pi - \frac{\gamma^\pi}{\epsilon} \right\|_\infty \le C. \tag{105}$$

No stationary attainment is needed in this step. Multiplying by ϵ and letting $\epsilon \downarrow 0$, Lemma D.1 identifies $\gamma^\pi = g^\pi$.

Step 3: optimal control. Every stationary policy has discounted value at most V_ϵ, and a single deterministic stationary discounted-optimal policy attains V_ϵ at all states, by Lemma C.1(ii)–(iii). Thus V_ϵ is the coordinatewise maximum of V_ϵ^π over $\pi \in \Pi_D$. Taking this finite maximum in the bounds from Step 2 yields

$$\left\| V_\epsilon - \frac{\bar{g}}{\epsilon} \right\|_\infty \le C, \qquad \bar{g}_i := \max_{\pi \in \Pi_D} g_i^\pi.$$

Theorem E.1 identifies $\bar{g} = g^\star$. The bounded centered family has a convergent subsequence as $\epsilon \downarrow 0$, and Lemma M.1 gives a finite optimal Bellman pair. For a fixed deterministic policy, Step 2 gives the bound directly, and the one-action version of Lemma M.1 gives the fixed-policy pair.

Step 4: randomized policies. Fix a stationary randomized π and define $\alpha_\pi := \min\{\pi(a \mid i) : \pi(a \mid i) > 0\} > 0$. If the jth coordinate of an effective row $\bar{p} := \sum_a \pi(a \mid i) p_a$ is positive, one of its nonnegative summands is positive, and that summand is at least $\alpha_\pi \delta$. Hence the effective row sets $\mathcal{U}_i^\pi$ are compact and have support gap $\alpha_\pi \delta$. Steps 1–2, applied to the one-action model with these effective row sets, give the bound on $V_\epsilon^\pi - g^\pi/\epsilon$ with a constant that may depend on π. The one-action version of Lemma M.1 on the rows $\mathcal{U}_i^\pi$ then gives a finite fixed-policy pair, and Step 2 of the proof of Proposition I.6 identifies its equations with equation 66–equation 67. □

Corollary M.5. *For every deterministic stationary controller π, define $\mathcal{P}^\pi := \{P^{\pi,q} : q \in \mathcal{Q}_S\}$. Suppose that $P \mapsto P^\infty$ is continuous on each compact family $\mathcal{P}^\pi$. Then the optimal discounted values have bounded vector centering as in equation 103, and the full vector Bellman system has a finite solution. The same conclusion holds for every fixed deterministic policy.*

Proof. **Step 1: bounded centering.** The proof of Lemma M.3 uses the support gap only to establish continuity of P^∞. Under the present hypothesis, the right side of equation 101 is continuous on each compact set $\mathcal{P}^\pi \times [0, 1]$ directly. Consequently, equation 104 holds with a uniform constant after maximizing over the finitely many π. Steps 2–3 of the proof of Theorem M.4 then give bounded vector centering and finite Bellman solvability.

Step 2: the stationary criterion. We also verify the three conditions of Theorem 5.2 directly. On the compact full-selector space $\mathcal{Q}_S$, every vector $\eta^{\pi,q} := (P^{\pi,q})^\infty r^\pi$ with $\pi \in \Pi_D$ is continuous in q, so $d_i(q) := \max_{\pi\in\Pi_D} \eta_i^{\pi,q}$ is continuous. Choose the selectors of equation 35 with errors tending to zero and take a convergent subsequence. Its limit $q^\star$ satisfies $d(q^\star) = g^\star$. Together with the common optimal controller $\pi^\star$, this gives the stationary gain guarantees of Theorem H.1.

Step 3: restriction to the tangent game. By Proposition M.10, $g^\star = \widehat{T} g^\star$, so the tangent objects $\Pi_{g^\star}$, $A_{g^\star}(i)$, and $F_{ia}(g^\star)$ are defined. To restrict the pair $(\pi^\star, q^\star)$ to the tangent game, we check activity, including nature's rows at unused active actions. By equation 31 for $\pi^\star$ and $g^{\pi^\star} = g^\star$, $m_{i,\pi^\star(i)}(g^\star) = g_i^\star$, so $\pi^\star \in \Pi_{g^\star}$. By Lemma C.6(ii) for $q^\star$ and $d(q^\star) = g^\star$,

$$\max_{a\in A(i)} (q_{ia}^\star)^\top g^\star = d_i(q^\star) = g_i^\star.$$

Thus $(q_{ia}^\star)^\top g^\star \le g_i^\star$ for every action. If $a \in A_{g^\star}(i)$, feasibility also gives $(q_{ia}^\star)^\top g^\star \ge m_{ia}(g^\star) = g_i^\star$. Equality holds, so every such row belongs to $F_{ia}(g^\star)$. Let $\bar{q}$ be the restriction of $q^\star$ to active actions, and set $\bar{\pi} := \pi^\star$. Then $\bar{q}$ specifies a feasible tangent row at every active action, including actions that $\bar{\pi}$ does not use.

Step 4: the three conditions. Every tangent pair preserves $g^\star$: $P^{\pi,q} g^\star = g^\star$, hence $(P^{\pi,q})^\infty g^\star = g^\star$. Its gain for rewards centered by $g^\star$ is therefore its original gain minus $g^\star$. A tangent selector can be extended by arbitrary feasible rows at inactive actions, without changing a tangent pair's transition matrix. The original stationary guarantees consequently imply equation 73–equation 74.

It remains to bound the canonical biases for zero-centered-gain replies to $\bar{\pi}$. Define $c_i := r_{i,\bar{\pi}(i)} - g_i^\star$. The relevant kernels form a compact subset of $\mathcal{P}^{\bar{\pi}}$. Continuity of P^∞ and invertibility of $I - P + P^\infty$ imply that $Z_P := (I - P + P^\infty)^{-1}$ is continuous and uniformly bounded on this subset. Whenever $P^\infty c = 0$, Lemma C.2 identifies the canonical bias as $Z_P c$. Its norm is thus uniformly bounded, which in particular gives the common componentwise lower bound in equation 75. □

Corollary M.6. *Suppose each compact row family $\mathcal{U}_{ia}$ has a fixed support: for each coordinate j, either $p_j = 0$ for every row in the family or $p_j > 0$ for every row. Then Theorem M.4 applies.*

Proof. For each (i, a, j) whose coordinate is positive throughout $\mathcal{U}_{ia}$, compactness and continuity give $\delta_{iaj} := \min_{p\in\mathcal{U}_{ia}} p_j > 0$. There are finitely many such triples, and each stochastic row family has at least one. Define $\delta := \min_{(i,a,j):\, \delta_{iaj} \text{ is defined}} \delta_{iaj} > 0$. Then

$$p_j = 0 \quad \text{or} \quad p_j \ge \delta_{iaj} \ge \delta \qquad (i, a,\ p \in \mathcal{U}_{ia},\ j),$$

which is equation 102. □

The support-gap condition allows several support patterns and any number of recurrent classes. Its one-player antecedent is Schweitzer (1985, Theorem 2). Here the uniform resolvent estimate and the finite maximization over controller policies give the robust optimality version. Compactness, convexity, and semialgebraicity alone imply neither the gap nor finite-bias solvability; the curved examples of Appendix M.2 exhibit the corresponding failure modes.

M.1.1 Divergence balls and the support-loss boundary

For a nominal row $p^0 \in \Delta(S)$, define $J(p^0) := \{j : p_j^0 > 0\}$ and $\Delta(J) := \{p \in \Delta(S) : p_j = 0 \text{ for } j \notin J\}$. We use $D_{\mathrm{KL}}(p\|p^0) := \sum_{j:p_j>0} p_j \log(p_j/p_j^0)$, with value $+\infty$ if $p_j > 0 = p_j^0$ for some j. The two orders of the KL divergence lead to different support conditions.

Corollary M.7. *Suppose each row family $\mathcal{U}_{ia}$ is either a polytope or is compact and satisfies $p_j = 0$ or $p_j \ge \delta_{ia}$ for all its rows and coordinates, where $\delta_{ia} > 0$. Then the full vector Bellman system has a finite solution $(g^\star, h)$, and $\sup_{0<\epsilon<1} \|V_\epsilon - g^\star/\epsilon\|_\infty < \infty$. The fixed-policy conclusions of Theorem M.4 also hold.*

Proof. **Step 1: a reduced model with a support gap.** Replace every polytopic $\mathcal{U}_{ia}$ by its finite vertex set E_{ia}. The minimum of $p^\top x$ is unchanged for every $x \in \mathbb{R}^S$, so this replacement leaves T, every fixed-policy operator T^π, and their discounted fixed points unchanged. The reduced row families are compact. They also have a common support gap: take the minimum of the gaps for the

unreplaced row families and all positive coordinates of the finitely many vertices. Only nonempty collections enter this minimum, and every vertex has at least one positive coordinate.

Step 2: bounded centering. Steps 1–3 of the proof of Theorem M.4, applied to the reduced model, give the bound on $V_\epsilon - g^\star/\epsilon$. These steps use compactness, not convexity, of the row families, and the reduced model has the same operator T and hence the same gain $g^\star$.

Step 3: the original row sets and fixed policies. Apply Lemma M.1 to the original discounted operator. The bound from Step 2 gives a finite Bellman pair for the original row sets. Equivalently, Lemma I.1 determines L_g from T, so the reduction preserves the tangent equations as well. For a fixed randomized policy, an effective row formed from the reduced families has a support gap equal to their common gap times the smallest positive action probability of that policy. The argument in Step 4 of Theorem M.4 therefore gives bounded centering for its unchanged operator T^π, and Lemma M.1 gives the fixed-policy pair. □

Corollary M.8. *For each (i,a), let p^0_{ia} be a nominal row, put $J_{ia} := J(p^0_{ia})$, and let $\rho_{ia} \geq 0$. Assume that every row family is one of the following:*

1. *a forward KL ball $\mathcal{U}_{ia} := \{p \in \Delta(S) : D_{\mathrm{KL}}(p\|p^0_{ia}) \leq \rho_{ia}\}$ with*
$$\rho_{ia} < \rho^{\mathrm{del}}_{ia} := \min_{j \in J_{ia}} -\log(1 - p^0_{ia,j}), \qquad -\log 0 := +\infty; \tag{106}$$
2. *a support-restricted reverse KL ball $\mathcal{U}_{ia} := \{p \in \Delta(J_{ia}) : D_{\mathrm{KL}}(p^0_{ia}\|p) \leq \rho_{ia}\}$ with $\rho_{ia} < \infty$; or*
3. *a polytope, including a total-variation ball.*

Then the conclusions of Corollary M.7 hold. For a full-support nominal row, the support restriction in item 2 is automatic.

Proof. By Corollary M.7, it suffices to show that every row family of type 1 or 2 is compact and has a fixed support. Corollary M.6 will then provide the required positive gap. Both balls contain their nominal row. They are closed because the corresponding extended KL divergence is lower semicontinuous, and they are compact because they are subsets of a finite-dimensional probability simplex.

Forward KL. Fix a forward KL row and abbreviate its nominal support by J. A finite forward divergence forces $p \in \Delta(J)$. If $|J| = 1$, the ball is the nominal singleton. Otherwise, for $j \in J$ and a row with $p_j = 0$, direct substitution and the nonnegativity of the KL divergence give

$$D_{\mathrm{KL}}(p\|p^0) = -\log(1 - p^0_j) + D_{\mathrm{KL}}\left(p \middle\| \frac{p^0|_{J\setminus\{j\}}}{1 - p^0_j}\right) \geq -\log(1 - p^0_j), \tag{107}$$

where the conditional distribution is extended by zero outside $J\setminus\{j\}$. Equality holds at this conditional distribution, so $-\log(1 - p^0_j)$ is the exact cost of deleting coordinate j. Under equation 106, no coordinate in J can vanish on the ball. Each such coordinate is continuous and positive on the compact ball, so it has a positive minimum there, and the support is exactly J.

Reverse KL. For a reverse KL row, $D_{\mathrm{KL}}(p^0\|p) < \infty$ forces $p_j > 0$ for each $j \in J$. More explicitly, with $H(p^0) := -\sum_{j\in J} p^0_j \log p^0_j$, and since every term $-p^0_k \log p_k$ is nonnegative,

$$D_{\mathrm{KL}}(p^0\|p) = -H(p^0) - \sum_{k\in J} p^0_k \log p_k \geq -H(p^0) - p^0_j \log p_j.$$

The inequality $D_{\mathrm{KL}}(p^0\|p) \leq \rho$ therefore implies $-p^0_j \log p_j \leq \rho + H(p^0)$, or $p_j \geq \exp[-(\rho + H(p^0))/p^0_j] > 0$. This bound holds for every $j \in J$ throughout the ball. Together with the restriction to $\Delta(J)$, it makes the support exactly J. □

The strict radius in equation 106 is the exact condition for a forward KL ball to preserve its nominal support. When the nominal support contains at least two states, the threshold is finite. At equality, the conditional nominal row from equation 107, for a minimizing coordinate j, lies in the ball and has

zero mass at j. Its mixtures with p^0 remain feasible by convexity and have arbitrarily small positive mass at j. Thus the support gap also fails at this finite threshold.

A larger forward radius may still give a solvable model through another condition. For example, if $\rho_{ia} \geq \log(1/\min_{j\in J_{ia}} p^0_{ia,j})$, then every $p \in \Delta(J_{ia})$ satisfies $D_{\mathrm{KL}}(p\|p^0_{ia}) \leq \log(1/\min_{j\in J_{ia}} p^0_{ia,j})$. The ball is then the entire simplex face and item 3 applies. If $|J_{ia}| \leq 2$, every forward KL ball is a closed interval or a singleton in that face, so it is a polytope at every radius. For reverse KL, finite divergence bounds the coordinates in J away from zero but does not bound positive coordinates outside J away from zero. This is why item 2 includes the support restriction when the nominal row is sparse.

The same reasoning applies to other divergences on a fixed nominal support. Let $f : [0,\infty) \to \mathbb{R} \cup \{+\infty\}$ be convex and lower semicontinuous, with $f(1) = 0$, and define $D_f(p\|p^0) := \sum_{j\in J} p^0_j f(p_j/p^0_j)$ on $\Delta(J)$. If $p_j = 0$ and $p^0_j < 1$, the remaining coordinates satisfy $\sum_{k\neq j} p_k = 1$. Apply Jensen's inequality with weights $p^0_k/(1-p^0_j)$ to obtain

$$D_f(p\|p^0) \geq \beta_f(p^0_j), \qquad \beta_f(t) := tf(0) + (1-t)f\left(\frac{1}{1-t}\right), \tag{108}$$

The bound is attained at $p = p^0(\cdot \mid J \setminus \{j\})$. Define $\beta_f(1) := +\infty$, since a probability row on a singleton support cannot delete its only coordinate. Thus a ball $\{p \in \Delta(J_{ia}) : D_f(p\|p^0_{ia}) \leq \rho_{ia}\}$ with $\rho_{ia} < \min_{j\in J_{ia}} \beta_f(p^0_{ia,j})$ has a fixed support and can replace either KL type in Corollary M.8. Its compactness follows from lower semicontinuity on the compact simplex face.

For $0 < t < 1$, the deletion costs are $-\log(1-t)$ for forward KL, $t/(1-t)$ for Pearson's $\chi^2(p\|p^0) := \sum_{j\in J}(p_j - p^0_j)^2/p^0_j$, and $1-\sqrt{1-t}$ for the squared Hellinger divergence $H^2(p,p^0) := 1 - \sum_j \sqrt{p_j p^0_j}$. For Hellinger balls with sparse nominal rows, the restriction to $\Delta(J)$ is part of this condition, because Hellinger divergence remains finite for rows charging coordinates outside J. Total-variation balls need no radius restriction for solvability because they are polytopes.

Example 5 (Failure at the first forward-KL support loss)**.** *There are three states (x, y, z) and one action per state. State y returns deterministically to x, state z is absorbing, and the rewards are $(r_x, r_y, r_z) = (1/2, -1, 0)$. At x, take the KL ball*

$$\mathcal{U}_x := \{p \in \Delta(\{x,y,z\}) : D_{\mathrm{KL}}(p\|(1/3,1/3,1/3)) \leq \log(3/2)\}.$$

Its radius equals equation 106. The robust gain is $g^\star = 0$, but the vector Bellman system has no finite solution.

Proof. **Step 1: compute the stationary gains.** Write a row as $p = (1-u-v, u, v)$, where u is the probability of moving to y and v is the probability of moving to z. On the face $v = 0$,

$$D_{\mathrm{KL}}(p\|(1/3,1/3,1/3)) = \log(3/2) + d_{\mathrm{Ber}}(u\|1/2),$$

Here $d_{\mathrm{Ber}}(u\|q) := u\log(u/q) + (1-u)\log((1-u)/(1-q))$, with the usual zero-term convention. This binary KL divergence vanishes only at $u = q$. Thus the only feasible row with $v = 0$ is $p^* := (1/2, 1/2, 0)$. Under p^*, the class $\{x, y\}$ has invariant distribution $(2/3, 1/3)$ and average reward $\frac{2}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot(-1) = 0$. State z has reward zero as well.

If a fixed feasible row has $v > 0$, each visit to x is followed by absorption at z with probability v. A failed attempt returns to x within two steps. Therefore absorption occurs almost surely and has finite expected time, at most $2/v$ from x. Since rewards after absorption are zero, this row has gain zero from every state. All stationary rows consequently have gain zero. Lemma D.1 identifies their coordinatewise infimum with the fixed-policy robust gain, and Theorem E.1 gives the same value against history-dependent nature. There is only one controller policy, so $g^\star = 0$.

Step 2: construct feasible rows with small positive leakage. For sufficiently small $t > 0$, define $p_t := (1/2 - t - t^2,\ 1/2 + t,\ t^2)$. Its coordinates are nonnegative and sum to one. To check feasibility, define $f(s) := s\log(3s)$. Taylor's formula at $s = 1/2$ gives $f'(1/2) = 1 + \log(3/2)$ and $f''(1/2) = 2$. Applying the formula to the first two coordinates and keeping the third term $f(t^2) = t^2\log(3t^2)$ exact yields

$$D_{\mathrm{KL}}(p_t\|(1/3,1/3,1/3)) - \log(3/2) = t^2\left[1 + \log(2t^2)\right] + O(t^3),$$

After division by t^2, the right side is $1 + \log(2t^2) + O(t)$, which tends to $-\infty$. Thus $p_t \in \mathcal{U}_x$ for all sufficiently small $t > 0$.

Step 3: rule out a finite bias. If a finite Bellman pair existed, Theorem I.3 would identify its gain as $g^\star = 0$. Every feasible row would then belong to the gain face. The bias equation at y would give $h_y = h_x - 1$. Substitution into the equation at x would give

$$0 = \min_{(1-u-v,u,v)\in\mathcal{U}_x} \{1/2 - u + v(h_z - h_x)\}.$$

For the feasible row p_t, the expression inside the minimum equals $-t + t^2(h_z - h_x)$. For any fixed finite h, it is negative when t is sufficiently small. This contradicts the required minimum of zero. The same obstruction appears in the canonical biases: the unique recurrent state of the p_t-chain is z, so the canonical normalization is $w_z = 0$. Its Poisson equations give $w_y = w_x - 1$ and $t^2 w_x = 1/2 - (1/2 + t) = -t$. Hence $w_x = -1/t$, which has no uniform lower bound as $t \downarrow 0$. □

M.2 Two distinct obstructions to finite Bellman solvability

The fixed-policy criterion separates two possible obstructions. Nature may have no stationary selector that attains its worst gain, or its gain-attaining selectors may have canonical biases with no common componentwise lower bound. The following models exhibit these mechanisms with state-dependent rewards and compact convex ambiguity sets described by polynomial inequalities. The same two mechanisms occur in the classical one-player examples of Schweitzer (1985, Examples 1–2).

Example 6 (Compact convex ambiguity without stationary attainment)**.** *There are states (s, m, z) with rewards $(0, -1, 0)$, and m, z are absorbing. At s, let*

$$\mathcal{U}_s := \{(1 - a, \beta, a - \beta) : 0 \le a \le 1,\ 0 \le \beta \le a(1 - a)\}. \tag{109}$$

This compact convex semialgebraic model has robust gain $(-1, -1, 0)$, but no stationary nature selector attains it from s, and no finite Bellman pair exists. Its ambiguity set and nonattainment calculation are those of Grand-Clement et al. (2023, Proposition 3.2); below we also identify the resulting Bellman obstruction.

Proof. **Step 1: check the ambiguity set.** The parameter set is closed and bounded. It is convex because the points below the graph of the concave function $a \mapsto a(1 - a)$ form a convex set, and the remaining constraints are linear. The affine map $(a, \beta) \mapsto (1 - a, \beta, a - \beta)$ therefore produces a compact convex set of probability rows. The description uses polynomial inequalities, so the set is semialgebraic.

Step 2: compute the gain and prove nonattainment. For a fixed row with $a > 0$, the process leaves s almost surely. Its probability of eventual absorption in m and its gain from s are

$$\mathbb{P}_s(\tau_m < \infty) = \sum_{t\ge 0} (1 - a)^t \beta = \beta/a, \qquad \eta_s = -\beta/a \ge -(1 - a) > -1.$$

At $a = 0$, the row keeps the process at s forever and has gain zero. Thus every stationary row has gain strictly larger than -1 from s. On the other hand, the feasible boundary rows with $\beta = a(1 - a)$ have gain $-(1 - a) \to -1$ as $a \downarrow 0$. Since every stage reward is at least -1, no history-dependent nature strategy can yield an average below -1 under any of the four payoff conventions. These two bounds show that the robust value from s is -1. The absorbing states have values -1 and 0, so the full gain is $(-1, -1, 0)$. No stationary row attains its s-coordinate.

Step 3: rule out a finite bias. By Theorem I.3, every finite Bellman pair must have this gain. For a feasible row,

$$(1 - a, \beta, a - \beta)^\top g = -1 + a - \beta \ge -1 + a^2.$$

Equality with $g_s = -1$ holds only at $a = \beta = 0$. The gain face is therefore the self-loop, and its bias equation $-1 + h_s = 0 + h_s$ is a contradiction. □

Example 7 (Stationary attainment with an unbounded lower bias envelope)**.** *Let the states be (x, y, z), with rewards $(0, -1, 0)$. State y returns to x, state z is absorbing, and*

$$\mathcal{U}_x := \mathrm{co}\{(1 - k - k^2, k, k^2) : 0 \le k \le 1/2\}.$$

The row set is compact, convex, and semialgebraic. Every stationary kernel has gain zero and attains the robust value, but the canonical biases have no common lower bound. No finite Bellman pair exists.

Proof. **Step 1: describe the convex hull.** Write a row as $(1-u-v, u, v)$. A convex combination with weights λ_j and parameters k_j satisfies

$$u = \sum_j \lambda_j k_j, \qquad v = \sum_j \lambda_j k_j^2 \geq u^2, \qquad v = 0 \iff u = 0,$$

where the inequality is Jensen's inequality. Compactness follows from the compact generating curve and Carathéodory's theorem. The parameter region is exactly $0 \leq u \leq 1/2$, $u^2 \leq v \leq u/2$. Necessity follows from the preceding inequality and $k^2 \leq k/2$ for $k \in [0, 1/2]$. Conversely, at a given u, the lower endpoint $v = u^2$ is generated by $k = u$, the upper endpoint $v = u/2$ by mixing $k = 0$ and $k = 1/2$, and their mixtures fill the interval. This also verifies semialgebraicity.

Step 2: compute the gains. If $u > 0$, then $v \geq u^2 > 0$. Under a fixed row, the probability of avoiding z for the first j visits to x is $(1-v)^j$. A return to x takes at most two steps, so absorption at z has finite expected time, at most $2/v$ from x. Hence z is the unique recurrent class and the gain is zero. If $u = v = 0$, the recurrent classes are $\{x\}$ and $\{z\}$, both with reward zero, and y incurs its reward only once before reaching x. This kernel also has gain zero from every state. Lemma D.1 identifies the infimum of these stationary gains with the fixed-policy gain. Theorem E.1 gives the same value against history-dependent nature. Thus the robust gain is zero.

Step 3: compute the canonical biases. For $u > 0$, the unique recurrent state is z, so zero mean on the recurrent class means $w_z = 0$. The Poisson equations give

$$w_y = w_x - 1, \qquad w_x = (1-u-v)w_x + uw_y = (1-v)w_x - u,$$

so $w_x = -u/v$ and $w_y = -1 - u/v$. Along the generating curve,

$$w(k) = (-1/k, -1 - 1/k, 0) \quad (k > 0), \qquad w(0) = (0, -1, 0).$$

All these kernels attain gain zero, but their first two bias coordinates tend to $-\infty$ as $k \downarrow 0$.

Step 4: rule out a finite Bellman bias. A finite bias at gain zero would satisfy $h_y = h_x - 1$. A linear function has the same minimum on the generating curve and its convex hull, so the equation at x would be

$$0 = \min_{0 \leq k \leq 1/2} \{-k + k^2(h_z - h_x)\}.$$

Define $H := h_z - h_x$. If $H \leq 0$, every positive k makes the expression negative. If $H > 0$, any $0 < k < \min\{1/2, 1/H\}$ does so, because $-k + k^2 H = k(-1 + kH) < 0$. Thus no such bias exists, and Theorem I.3 excludes a pair with a different gain. □

Corollary J.2 diagnoses the two examples clause by clause: the first violates simultaneous worst-gain attainment, and the second violates the one-sided canonical-bias envelope. Their statewise average values nevertheless remain well defined by Appendix E, which establishes the value without assuming a finite bias.

M.3 Representation invariance and elementary stability

Proposition M.9. *The following statements hold without restrictions on recurrent classes.*

(i) Replacing each $\mathcal{U}_{ia}$ by $\mathrm{co}(\mathcal{U}_{ia})$ leaves the discounted, finite-horizon, and robust average values unchanged, both for optimal control and for every fixed stationary randomized controller. It also leaves the set of average-optimal stationary controllers unchanged.

(ii) With the ambiguity sets fixed, write $g^\star(r)$ for the optimal gain under the reward array r. If $\delta := \max_{i,a} |r_{ia} - \widetilde{r}_{ia}|$, then

$$\|g^\star(r) - g^\star(\widetilde{r})\|_\infty \leq \delta.$$

The same bound holds for every fixed-policy gain. The gain is monotone in the rewards, satisfies $g^\star(r + c) = g^\star(r) + c\mathbf{1}$ for every scalar reward shift c, and satisfies $g^\star(\alpha r) = \alpha g^\star(r)$ for $\alpha \geq 0$.

(iii) If $\mathcal{U}_{ia} \subseteq \widetilde{\mathcal{U}}_{ia}$ at every pair, then $g^\star(r,\widetilde{\mathcal{U}}) \le g^\star(r,\mathcal{U})$. The same order holds for every fixed-policy gain.

Proof. **(i)** Linear minimization over a set and over its convex hull gives the same value. In finite dimension, $\mathrm{co}(\mathcal{U}_{ia})$ is compact, so convexification leaves T, every T^π, and their discounted and finite-horizon values unchanged. Their normalized limits, and hence the gains, are unchanged by Theorems 3.1 and 3.2. A stationary policy is all-state average optimal exactly when $g^\pi = g^\star$, so the set of such policies is unchanged as well.

(ii) For rewards at distance δ in the supremum norm, $T_{\tilde{r}}x - \delta\mathbf{1} \le T_r x \le T_{\tilde{r}}x + \delta\mathbf{1}$. Monotonicity and additive homogeneity give, by induction on N,

$$T_{\tilde{r}}^N 0 - N\delta\mathbf{1} \le T_r^N 0 \le T_{\tilde{r}}^N 0 + N\delta\mathbf{1}.$$

Dividing by N and taking the finite-horizon limits of Theorem 3.2 proves the Lipschitz bound. The same comparison without an error term proves monotonicity. The identities $T_{r+c}^N 0 = T_r^N 0 + Nc\mathbf{1}$ and $T_{\alpha r}^N 0 = \alpha T_r^N 0$ for $\alpha \ge 0$ give the scalar-shift and positive-homogeneity properties after normalization.

(iii) Larger ambiguity sets decrease every row minimum. Monotonicity propagates this operator order through all finite-horizon iterates, and the normalized limits give the gain order. Each argument applies unchanged to T^π, whose outer weights are nonnegative and sum to one. □

Proposition M.10. *The optimal gain satisfies $g^\star = \widehat{T} g^\star$. For every fixed stationary randomized controller,*

$$g_i^\pi = \sum_a \pi(a \mid i) \min_{p \in \mathcal{U}_{ia}} p^\top g^\pi.$$

These gain-only equations do not determine the average reward.

Proof. Define $v_\epsilon := \epsilon V_\epsilon$, which converges to $g^\star$ by Theorem 3.2. The discounted equation gives $(v_\epsilon)_i = \max_a\{\epsilon r_{ia} + (1-\epsilon) m_{ia}(v_\epsilon)\}$, and since $|p^\top x| \le \|x\|_\infty$ for probability rows,

$$\|v_\epsilon - \widehat{T} g^\star\|_\infty \le \epsilon R + (1-\epsilon)\|v_\epsilon - g^\star\|_\infty + \epsilon\|g^\star\|_\infty \longrightarrow 0.$$

This identifies the limit as $g^\star = \widehat{T} g^\star$. The fixed-policy equation is equation 31; the same estimate, with the action maximum replaced by the π-weighted sum, proves it. Every constant vector $c\mathbf{1}$ satisfies either recession equation, whatever the rewards. In a one-state model with reward r, however, the average gain is r. Thus the recession equation alone cannot identify the reward-dependent gain. □