STRATEGY COMPLETENESS AS A STATE-ABSTRACTION OBJECTIVE FOR MULTI-AGENT MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies state abstraction in multi-agent Markov games through the lens of strategy completeness—the requirement that a compressed policy class preserve the capacity to best-respond against strategic opponents. We show that on-policy predictive fidelity and fixed-profile value calibration are insufficient in competitive environments: an abstract representation can achieve zero transition prediction error and exact value calibration while incurring linear best-response loss against benchmark opponents. To explain this gap, we establish an exponential separation between decision preservation and value reconstruction, proving that preserving optimal responses against an opponent family can require exponentially fewer abstract states than reconstructing the underlying values of the same tasks under arbitrary action-dependent transitions. Motivated by this separation, we formulate the joint optimization of state partitions and policy profiles as a factorized Bellman program. We develop a provably efficient algorithm that exploits bounded treewidth in the primal graph induced by the joint Bellman factorization, computing an ε-optimal abstraction with certified lower and upper bounds in time polynomial in the explicit input size, H, and $1/\varepsilon$ for fixed treewidth. Finally, experiments on Markov Soccer and Connect Four validate the theory, showing that best-response-oriented state abstraction reduces best-response loss compared to predictive and value-based baselines under matched representation budgets.

1 INTRODUCTION

State abstraction is a foundational framework for scaling reinforcement learning to large state spaces by mapping environment states to a compact representation. Classical aggregation and bisimulation metrics preserve transition dynamics and reward structures, guaranteeing near-optimal planning in single-agent Markov decision processes (MDPs) (Givan et al., 2003; Li et al., 2006; Ferns et al., 2004; Abel et al., 2016). More recently, value-equivalence principles have relaxed full-model preservation by retaining only the Bellman evaluations necessary for a chosen value function class (Grimm et al., 2020; Ayoub et al., 2020). In multi-agent systems, abstraction is traditionally evaluated by equilibrium transfer—measuring whether an equilibrium computed in a reduced game remains an approximate equilibrium when lifted back to the original game (Sandholm & Singh, 2012; Ishibashi et al., 2025).

A key distinction between single-agent and multi-agent systems is that performance in Markov games depends intrinsically on the policies of other agents. In PSRO, league training, and competitive deployment, the downstream task continually shifts as new opponents are encountered. An abstract representation can perfectly predict on-policy trajectory distributions or preserve values for a fixed profile, yet entirely discard the state distinctions required to formulate an effective counter-strategy. Conversely, enforcing full model fidelity or reconstructing arbitrary state-action values retains substantial environmental variation that has no bearing on strategic action selection. This gap motivates examining *strategy completeness*—the property that a restricted policy class contains an exact best response (Zhang et al., 2023; Cui & Du, 2022b)—as a direct, constructive objective for state abstraction.

A Motivating Inadequacy of Prediction and Value Calibration. To illustrate the failure of predictive and value-based criteria in games, consider a two-state discounted game with states $\{s_0, s_1\}$

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.

sampled uniformly at each step, and a focal player choosing among actions $\{\text{safe}, \text{left}, \text{right}\}$:

	safe	left	right
s_0	$1/2$	1	0
s_1	$1/2$	0	$1.$

Suppose an encoder compresses $\{s_0, s_1\}$ into a single abstract state. Under the baseline safe policy, the reward is deterministically $1/2$ and the transition distribution is uniform regardless of state. Hence, on-policy model prediction is exact, and all abstract policies have expected return $H/2$ from the uniform initial distribution, where $H = (1-\gamma)^{-1}$. However, in this induced response MDP, an unrestricted response chooses left in s_0 and right in s_1, securing return H. The safe policy attains $H/2$ at both states, and no single-state policy improves its worst-state value, so $S_q(\phi) = H/2$. Similarly, an exact abstract equilibrium can coexist with an $H/2$ best-response loss against a fixed opponent (Appendix D). Thus, zero predictive error, value calibration, and equilibrium preservation fail to ensure strategy completeness.

Main Contributions. In this work, we formalize strategy completeness as a quantitative state-abstraction criterion in multi-agent Markov games. Our main contributions are:

1. **Strategic Abstraction Framework.** We define the best-response loss $S_q(\phi)$ for an opponent profile q under a state partition ϕ, and formulate the joint abstraction objective $J_K(\mathcal{G}, \mathcal{Q})$, which simultaneously optimizes a shared state partition of budget K, a deployed equilibrium profile, and task-specific response policies.
2. **Exponential Separation between Decision Preservation and Value Reconstruction.** For arbitrary action-dependent transitions, we prove an exponential separation between decision preservation and value reconstruction: achieving tolerance ε_0 on a benchmark opponent family requires $2^{|T|}$ abstract states for strategy completeness, whereas reconstructing the optimal values of the same tasks requires 2^{m+n_0} states (Theorem 4.1).
3. **Provably Efficient Joint Optimization via Bounded Treewidth.** We show that the non-convex joint Bellman formulation can be cast onto a factor graph. When the primal graph induced by the resulting finite factorization has bounded treewidth w, we develop an exact discretization and min–max dynamic programming algorithm that returns a certified ε-optimal abstraction in time polynomial in the explicit input size, H, and $1/\varepsilon$ for fixed w (Theorem 5.1).
4. **Scalable Certification and Empirical Validation.** We establish a spatial decay bound for local response certification in networked games and derive a unified error budget for PSRO. On Markov Soccer and a 161,029-state Connect Four benchmark, we show that optimizing best-response preservation lowers best-response loss compared to predictive and value-based baselines under matched state budgets.

2 Related Work

State Abstraction in MDPs and Value Equivalence. State aggregation and bisimulation metrics merge states based on reward and transition equivalence (Givan et al., 2003; Li et al., 2006; Ferns et al., 2011; Abel et al., 2016). Classical model minimization, homomorphisms, and approximate aggregation provide closely related quotient constructions and error guarantees (Dean & Givan, 1997; Dean et al., 1997; Ravindran & Barto, 2003; Ferns et al., 2004) (Hutter, 2014; Jiang et al., 2015; Ortner et al., 2014; Saldi et al., 2017). Latent representation methods similarly learn embeddings that preserve transition structure or behavioral similarity (Gelada et al., 2019; François-Lavet et al., 2019; Zhang et al., 2021; Castro et al., 2021); rich-observation theory learns latent decoders or low-rank features (Du et al., 2019; Misra et al., 2020; Agarwal et al., 2020; van der Pol et al., 2020). Predictive-state and successor representations summarize future observations or occupancies (Littman et al., 2001; Singh et al., 2004; Dayan, 1993; Barreto et al., 2017), with successor features extending reuse across tasks (Barreto et al., 2018). While reusable, these models retain task-irrelevant environmental complexity. Value-equivalence approaches (Grimm et al., 2020; Ayoub et al., 2020) and value-aware model losses (Farahmand et al., 2017) address this by preserving Bellman evaluations for fixed policy and probe classes. Multi-task abstraction makes a related trade-off across task distributions (Abel et al., 2018). Our framework differs in its preservation target: rather than preserving the numerical values of a probe class, we preserve the action optimality of best responses against an opponent family.

Table 1: Comparison of state-abstraction paradigms in Markov games.

Abstraction Paradigm	Mathematical Target	Preserved Invariant	Vulnerability in Multi-Agent Games
Model / Bisimulation (Givan et al., 2003; Abel et al., 2016)	Transitions and rewards: $P(\cdot \mid s, a), r(s, a)$	Environment dynamics across all actions	Encodes state distinctions irrelevant to competitive response
Value Equivalence (Grimm et al., 2020; Ayoub et al., 2020)	Bellman evaluation: $\mathcal{T}^{\pi} V$ for $\pi \in \Pi_{\text{probe}}$	Numerical values of specified probe families	Retains value-shifting variations that do not affect optimal action selection
Equilibrium Transfer (Sandholm & Singh, 2012; Ishibashi et al., 2025)	Lifted equilibrium profile: $\text{Gap}_{\mathcal{G}}(\pi)$	Original-game unilateral deviation gains	May omit best responses to other opponents
Strategy Completeness (Ours)	Best-response operator: $\mathcal{T}^{*,q} W_i^q$ for $q \in \mathcal{Q}$	Optimal counter-actions against opponent pool $\mathcal{Q}$	Requires coverage of the intended opponent family

Game Abstraction and Equilibrium Transfer. Extensive-form and stochastic game abstractions aggregate game states to compute abstract equilibria that lift with bounded regret (Gilpin & Sandholm, 2007b; Kroer & Sandholm, 2014; Sandholm & Singh, 2012; Ishibashi et al., 2025). Automated poker abstraction provides early large-scale examples (Gilpin & Sandholm, 2006; 2007a; Gilpin et al., 2007), while imperfect-recall refinement and bilinear solution methods study related abstraction–optimization trade-offs (Čermák et al., 2017; Bošanský et al., 2017). These methods focus on deployment fidelity (E_ϕ). In contrast, strategy-complete abstraction addresses the dual requirement: ensuring that the compressed representation retains the strategic flexibility to respond to multiple distinct opponents ($S_{\mathcal{Q}}$).

Strategy Completeness in Offline MARL. In offline Markov game analyses, strategy completeness and unilateral coverage conditions are necessary to prevent out-of-distribution exploitation (Cui & Du, 2022b;a; Zhang et al., 2023; Freihaut et al., 2025) (Freihaut et al., 2026; Bergerault et al., 2026; Viano et al., 2026; Zhan et al., 2025). While prior work treats completeness as a static mathematical assumption on a pre-given policy class, our work operationalizes it as a constructive objective to synthesize optimal state abstractions under explicit representation constraints. Population methods determine which opponents enter the task set (Lanctot et al., 2017; Balduzzi et al., 2019; Vinyals et al., 2019); localized MARL provides complementary scaling mechanisms when interaction influence decays spatially (Qu et al., 2022; Zhou et al., 2023; Ren et al., 2025; Chakraborty et al., 2026).

3 Strategy Completeness as a State-Abstraction Objective

We consider a finite discounted n-player Markov game $\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{r_i\}_{i\in\mathcal{N}}, \gamma)$ (Shapley, 1953; Fink, 1964), where $\mathcal{N} = \{1, \ldots, n\}$ is the set of players, $\mathcal{S}$ is the finite state space, $\mathcal{A}_i$ is the finite action space of player i with joint action space $\mathcal{A} = \prod_{i=1}^n \mathcal{A}_i$, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is player i's reward function, and $\gamma \in [0, 1)$ is the discount factor with effective horizon $H := (1 - \gamma)^{-1}$. Let $\Pi_i = \mathcal{S} \to \Delta(\mathcal{A}_i)$ denote stationary Markov policies for player i, and $\Pi = \prod_{i=1}^n \Pi_i$. We denote opponents of player i by $-i$, such that $q \in \Pi_{-i}$ represents a joint stationary opponent strategy.

For a fixed opponent profile $q \in \Pi_{-i}$, player i's induced environment is a single-agent MDP with Bellman evaluation operator $\mathcal{T}_i^{\pi_i,q} v := r_i^{\pi_i,q} + \gamma P^{\pi_i,q} v$ and optimality operator $\mathcal{T}_i^{*,q} v := \max_{a_i \in \mathcal{A}_i} [r_i(\cdot, a_i, q) + \gamma P(\cdot, a_i, q) v]$. The unrestricted best-response value is the unique fixed point $W_i^q = \mathcal{T}_i^{*,q} W_i^q$.

Definition 3.1 (State Abstraction and Abstract Policy Class)**.** *A state abstraction with cardinality budget K is a partition mapping $\phi : \mathcal{S} \to [K]$ using at most K labels. The abstraction induces a restricted class of stationary policies constant on each partition cell:*

$$\Pi_\phi := \{\pi \in \Pi \mid \forall s, s' \in \mathcal{S},\ \phi(s) = \phi(s') \implies \pi(\cdot \mid s) = \pi(\cdot \mid s')\}. \tag{1}$$

Definition 3.2 (Best-Response Loss). *Let $\mathcal{Q} \subset \Pi_{-i}$ be a finite benchmark set of opponent profiles. For any opponent $q \in \mathcal{Q}$, the best-response loss of abstraction ϕ for player i is*

$$S_q(\phi) := \inf_{\sigma \in \Pi_{\phi,i}} \|W_i^q - V_i^{\sigma,q}\|_\infty. \tag{2}$$

The abstraction loss across benchmark set $\mathcal{Q}$ under budget K is

$$D_K(\mathcal{G}, \mathcal{Q}) := \min_{\phi:|\phi(\mathcal{S})|\le K} \max_{q\in\mathcal{Q}} S_q(\phi). \tag{3}$$

Write $S_{\mathcal{Q}}(\phi) = \max_{q\in\mathcal{Q}} S_q(\phi)$. The focal policy class $\Pi_{\phi,i}$ is *strategy-complete* with respect to q if and only if $S_q(\phi) = 0$. When several focal players are benchmarked simultaneously, one may equivalently index response tasks by $h = (i, q)$, where i is the focal player and $q \in \Pi_{-i}$ is the opponent profile; we use the simpler notation above when the focal player is fixed.

Joint Abstraction Objective. In competitive games, an agent must simultaneously execute a low-Nash-gap deployment profile while retaining internal best-response representations for counter-strategies. Let $\mathrm{Gap}_{\mathcal{G}}(\pi) := \max_{i\in\mathcal{N}} \|W_i^{\pi_{-i}} - V_i^{\pi}\|_\infty$ denote the worst-player, worst-state Nash gap. We define the equilibrium-transfer error of partition ϕ as $E_\phi(\mathcal{G}) := \inf_{\pi\in\Pi_\phi} \mathrm{Gap}_{\mathcal{G}}(\pi)$, and formulate the joint objective:

$$J_K(\mathcal{G}, \mathcal{Q}) := \min_{\phi:|\phi(\mathcal{S})|\le K} \max\left\{ E_\phi(\mathcal{G}),\ \max_{q\in\mathcal{Q}} S_q(\phi) \right\}. \tag{4}$$

Here $i(q)$ denotes the focal player of task q (equal to i for a fixed-player family). Equivalently, J_K optimizes a single shared partition ϕ, a deployed profile $\pi \in \Pi_\phi$, and an opponent-specific response policy $\sigma_q \in \Pi_{\phi,i(q)}$ for each $q \in \mathcal{Q}$.

Figure 1 relates the three preservation targets to the shared partition and its deployment and response policies.

Exact Joint Bellman Characterization. For a candidate tuple $(\phi, \pi, \{\sigma_q\})$, let v_i, u_i and v_q, u_q denote state evaluation and supersolution witnesses. The exact program minimizes $t \ge 0$ subject to:

$$\begin{aligned} &v_i = \mathcal{T}_i^{\pi} v_i, && u_i \ge \mathcal{T}_i^{a_i,\pi_{-i}} u_i \ (\forall a_i \in \mathcal{A}_i), && u_i - v_i \le t\mathbf{1}, \\ &v_q = \mathcal{T}_i^{\sigma_q,q} v_q, && u_q \ge \mathcal{T}_i^{a_i,q} u_q \ (\forall a_i \in \mathcal{A}_i), && u_q - v_q \le t\mathbf{1}, \end{aligned} \tag{5}$$

where π and σ_q are constrained to be ϕ-abstract. In Appendix C, we prove that the global minimum of program (5) equals $J_K(\mathcal{G}, \mathcal{Q})$.

4 Exponential Separation between Decision Preservation and Value Reconstruction

While predictive and value-based representations retain sufficient detail to reconstruct numerical values, strategy completeness requires only identifying optimal decision boundaries. The following theorem demonstrates that this distinction can lead to an exponential gap in representation complexity, even under identical benchmark opponent tasks and arbitrary action-dependent dynamics.

Theorem 4.1 (Separation between Decision Preservation and Value Reconstruction). *Fix $m \ge 1, n_0 \ge 0, \gamma \in (0,1)$, and let P be an arbitrary joint-action transition kernel on state space $\mathcal{S} = \{0,1\}^{m+n_0}$. There exist bounded two-player constant-sum stage rewards and a benchmark opponent family $\mathcal{Q}_T$ indexed by a non-empty subset $T \subseteq [m]$ such that, at a fixed accuracy tolerance $\varepsilon_0 = 1/64$:*

$$K_{\text{equilibrium}} = 1, \qquad K_{\text{BR}}(\mathcal{Q}_T) = 2^{|T|}, \qquad K_{\text{value}}(\mathcal{Q}_T) = 2^{m+n_0}. \tag{6}$$

Here $K_{\text{BR}}(\mathcal{Q}_T)$ is the minimum partition cardinality required to achieve $S_{\mathcal{Q}_T}(\phi) \le \varepsilon_0$, while $K_{\text{value}}(\mathcal{Q}_T)$ is the minimum cardinality required by any encoder from which the optimal values W^q of the same tasks can be uniformly reconstructed within error ε_0. Furthermore, the best-response and value-reconstruction separation is robust to small uniform perturbations of transitions and rewards.

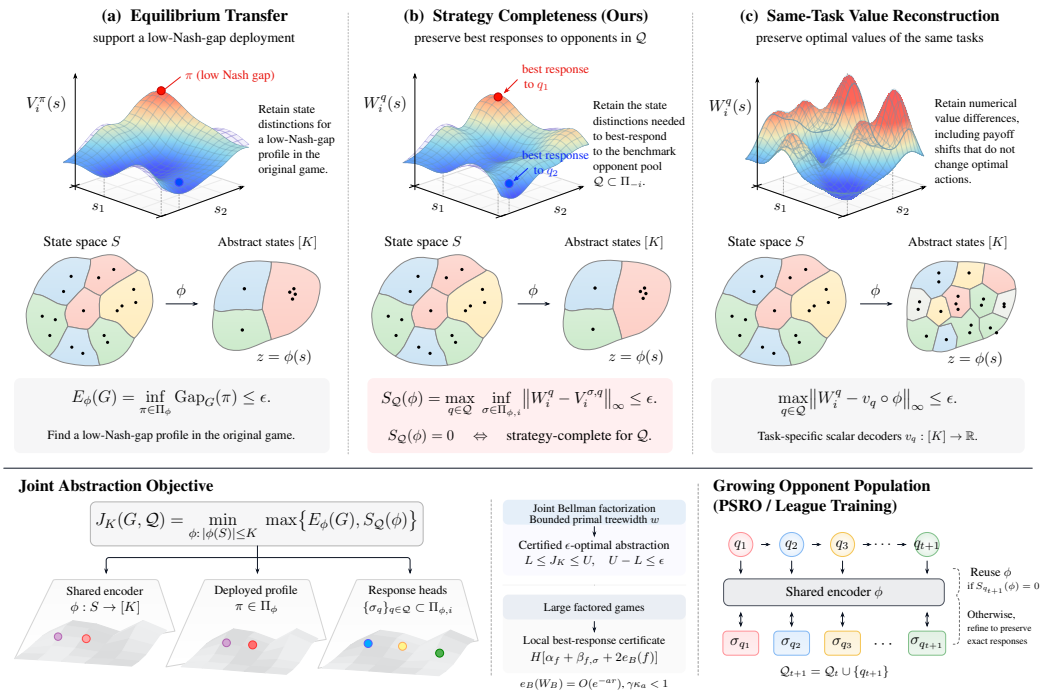

Figure 1: Preservation targets and joint state abstraction. A shared partition supports a low-Nash-gap deployment and best responses to a benchmark opponent family. Value reconstruction additionally retains numerical value differences. Global optimization and local certification provide complementary guarantees; exact response preservation under population growth may require a larger partition budget. The surfaces and partitions are schematic. Spatial decay assumes an exact local optimal critic and the conditions in Proposition 6.1.

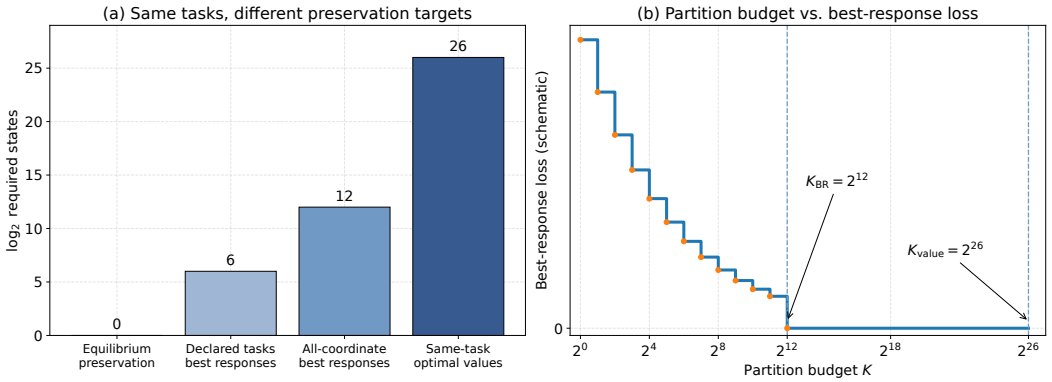

Figure 2: Exponential separation of preservation targets, with $m = 12$, $n_0 = 14$, and $\varepsilon_0 = 1/64$. (a) Required partition cardinalities: equilibrium, declared tasks with $|T| = 6$, all-coordinate responses, and same-task values. (b) Schematic loss trend for $T = [m]$ as partition budget K increases; exact responses are attainable at 2^{12} states, while same-task value reconstruction requires 2^{26}. Thresholds follow Theorem 4.1; the staircase illustrates the qualitative trend.

Proof Sketch. We decompose each state into policy-relevant coordinates $x \in \{0,1\}^m$ and payoff-shifting coordinates $y \in \{0,1\}^{n_0}$, setting $z = (x, y)$. For each task $q = (j, k, c) \in T \times [m + n_0] \times \{0, 1\}$, the reward structure is designed such that the optimal action is uniquely determined by coordinate x_j via $a_q^*(s) = x_j \oplus c$, yielding a strictly positive one-step Bellman deficit $g_q(s) \ge 1/8$ for the sub-optimal action. Consequently, achieving $S_q(\phi) \le \varepsilon_0$ requires separating states with distinct

Algorithm 1 Treewidth-Bounded Joint Abstraction and Policy Optimization

1: **Input:** Markov game $\mathcal{G}$, opponent pool $\mathcal{Q}$, partition budget K, target tolerance $\varepsilon > 0$, tree decomposition $\mathcal{T}$ of the primal graph induced by $\mathcal{G}_{\mathrm{B}}$ with treewidth w.
2: **Discretization:** Set resolution $M = \lceil H^2(2w+5)/\varepsilon \rceil$ and Bellman slack $\beta_M = (d_A+1)/M$, where $d_A = \sum_{i\in\mathcal{N}}(|\mathcal{A}_i|-1)$. Discretize all continuous probability and normalized value variables onto grid $\mathcal{G}_M = \{0, 1/M, \ldots, 1\}$.
3: **Factor Table Construction:** For each bag in $\mathcal{T}$, construct local min–max factor tables incorporating exact selector constraints and β_M-relaxed Bellman inequalities.
4: **Min–Max Elimination:** Traverse tree decomposition $\mathcal{T}$ from leaves to root, eliminating local variables via min–max dynamic programming messages:

$$m_r(x_{B_r}) = \min_{x_{v_r}\in\mathcal{X}_{v_r}} \max_{f\in\mathcal{B}_r} f(x_{\mathrm{scope}(f)}). \tag{7}$$

5: **Backtracking and Decoding:** Perform reverse-order backtracking to extract optimal discrete state partition $\widehat{\phi} : \mathcal{S} \to [K]$, deployed profile $\widehat{\pi} \in \Pi_{\widehat{\phi}}$, and response heads $\widehat{\sigma}_q \in \Pi_{\widehat{\phi},i(q)}$ $(q \in \mathcal{Q})$.
6: **Output:** System tuple $(\widehat{\phi}, \widehat{\pi}, \{\widehat{\sigma}_q\})$, certified lower bound $L = [Hc_M - H/M]_+$, and upper bound $U = \min\{H, Hc_M + 2H^2\beta_M\}$ satisfying $U - L \le \varepsilon$.

projections on T, which requires exactly $2^{|T|}$ abstract states. In contrast, the optimal value function satisfies $W^q(s) = H(1/2 + w_j) + \nu z_k$ with $\nu = 1/16$. Any two states differing in coordinate k have an optimal value discrepancy $|W^q(s) - W^q(s')| \ge 1/16 > 2\varepsilon_0$. Reconstructing task values therefore requires separating every pair of states in $\mathcal{S}$, demanding 2^{m+n_0} states. The detailed construction and perturbation bounds are provided in Appendix D; the zero-loss specialization also yields a graph-coloring characterization and a 3-colorability reduction (Karp, 1972).

5 Global Optimization via Bounded-Treewidth Dynamic Programming

The joint abstraction problem (4) is non-convex due to the coupling of discrete partition variables $\phi(s) \in [K]$, continuous action probabilities, and Bellman continuation values. Rather than restricting environment transition dynamics, we exploit the graph-theoretic structure of the *optimization problem*, following the use of graphical structure and variable elimination in games and optimization (Kearns et al., 2001; Dechter, 1999; Bienstock & Muñoz, 2015; Faenza et al., 2018).

We construct the joint Bellman factorization and its associated factor graph $\mathcal{G}_{\mathrm{B}}$: its variable nodes consist of state assignments $z_s \in [K]$, abstract action probabilities $x_{k,g,a}$, local probabilities $p_{s,g,a}$, and normalized value witnesses $v_g(s), u_g(s) \in [0,1]$; factor nodes enforce simplex normalization, selector constraints ($z_s = k \implies p_{s,g} = x_{k,g}$), and Bellman inequalities.

Theorem 5.1 (Global Bounded-Treewidth Optimization)**.** *For every finite discounted Markov game $\mathcal{G}$, opponent benchmark pool $\mathcal{Q}$, partition budget K, and tolerance $0 < \varepsilon \le H$, discretize continuous variables on a denominator-M grid with $M = \lceil H^2(2w+5)/\varepsilon \rceil$. If the primal graph induced by $\mathcal{G}_{\mathrm{B}}$ has treewidth w, Algorithm 1 returns an abstract partition $\widehat{\phi}$, policy profile $\widehat{\pi} \in \Pi_{\widehat{\phi}}$, response heads $\widehat{\sigma}_q \in \Pi_{\widehat{\phi},i(q)}$ $(q \in \mathcal{Q})$, and a certificate $[L, U]$ such that:*

$$L \le J_K(\mathcal{G}, \mathcal{Q}) \le F_{\mathcal{G}}(\widehat{f}) \le U \le L + \varepsilon, \tag{8}$$

where $F_{\mathcal{G}}(\widehat{f}) := \max\{\mathrm{Gap}_{\mathcal{G}}(\widehat{\pi}), \max_{q\in\mathcal{Q}} \|W^q_{i(q)} - V^{\widehat{\sigma}_q,q}_{i(q)}\|_\infty\}$. The computational complexity is bounded by

$$\mathrm{poly}(\mathcal{I}, \log M)\max\{K, M+1\}^{w+1}, \tag{9}$$

where $\mathcal{I}$ denotes the explicit input size. In particular, for fixed treewidth w, the running time is polynomial in $\mathcal{I}$, H, and $1/\varepsilon$. The guarantee holds for arbitrary action-dependent transitions.

Proof Overview. The proof proceeds in four steps. First, we round an exact continuous optimum to the grid $\mathcal{G}_M$. By Bellman contraction, shifting continuation values by $1/(2M)$ and policy probabilities

by d_A/M inflates Bellman residuals by at most $\beta_M = (d_A + 1)/M$, establishing that the discretized min–max optimum c_M provides a valid lower bound $L = [Hc_M - H/M]_+ \leq J_K$. Second, because selectors and probability simplexes are represented exactly on the grid, decoding the discrete assignment yields an admissible K-partition policy system $\widehat{f}$. Applying approximate Bellman witness bounds (Lemma E.2) shows that the true evaluated loss satisfies $F_{\mathcal{G}}(\widehat{f}) \leq U := \min\{H, Hc_M + 2H^2\beta_M\}$. Third, the gap satisfies $U - L \leq H^2(2w + 5)/M \leq \varepsilon$ by our choice of M. Finally, min–max variable elimination over a tree decomposition of width w exactly solves the discretized program in time $\mathrm{poly}(\mathcal{I}, \log M)D_M^{w+1}$ (Appendix E). For high-width instances, Appendix F uses McCormick envelopes (McCormick, 1976) in a mixed-integer outer relaxation, complementing related bilinear branch-and-bound methods in game abstraction (Bošanský et al., 2017). Appendix G transfers the resulting interval through a same-data concentration event; its reward-only specialization uses empirical Bernstein bounds (Maurer & Pontil, 2009).

6 Scalable Extensions: Networked Games and Population-Based MARL

6.1 Local Certification for Large Factored Games

When the induced primal graph has high treewidth, solving the full global program becomes computationally demanding. In large networked systems, we establish a local certification alternative by bounding boundary truncation error.

Let the state space factor as $\mathcal{S} = \prod_{j=1}^{n} \mathcal{S}_j$. For a focal agent, freeze opponent strategies, select a local subsystem of coordinates $B \subset [n]$, and construct a truncated local MDP on $\mathcal{S}_B$ with reference boundary s^0_{-B}. For a local critic $f : \mathcal{S}_B \to \mathbb{R}$ and local response σ, let α_f and $\beta_{f,\sigma}$ denote local Bellman residuals, and let $C = E^s + E^a\Pi^{(-i)}$ represent the induced-kernel sensitivity matrix (Appendix I).

Proposition 6.1 (Local Best-Response Certificate)**.** *Let* $e_B(f) = \sum_{j \notin B} b_j + \gamma \sum_{k \in B} \delta_k(f) \sum_{j \notin B} C_{kj}$ *bound the influence crossing the boundary. The lifted local policy* σ *satisfies:*

$$0 \leq W_i^q(s) - V_i^{\sigma,q}(s) \leq H\left[\alpha_f + \beta_{f,\sigma} + 2e_B(f)\right]. \tag{10}$$

For the exact local optimal critic $f = W_B$ *on* $B = B_r(i)$*, let* $\kappa_a = \max_k \sum_j C_{kj} e^{ad(k,j)}$ *and* $B_a = \sum_j b_j e^{ad(i,j)}$*. If* $\gamma\kappa_a < 1$*, then* $e_B(W_B) \leq B_a e^{-ar}/(1 - \gamma\kappa_a)$*. A network-size-independent bound also requires uniform control of* B_a *and the contraction margin.*

Proposition 6.1 supplies a computationally efficient upper certificate for large factored games without requiring global tree decompositions.

6.2 Interface with PSRO and League Training

In PSRO and league training, the benchmark opponent set $\mathcal{Q}$ expands iteratively. Strategy-complete abstraction provides an operational interface: adding an opponent appends a response task. Existing lower bounds remain valid; an upper bound for the enlarged pool requires evaluating the new response head (Appendix H).

Proposition 6.2 (Population Deviation Guarantee)**.** *Suppose the returned profile* $\pi \in \Pi_\phi$ *has worst-state deviation gain at most* τ *over the entire abstract policy class, up to the operational errors below. For each player* i*, assume its deployed opponents* $q = \pi_{-i}$ *belong to the benchmark task family. If abstraction* ϕ *satisfies best-response loss* $S_{\mathcal{Q}}(\phi) \leq \varepsilon_{\mathrm{BR}}$*, and planning, statistical, and numerical optimization errors are bounded by* $\varepsilon_{\mathrm{plan}}$*,* $\varepsilon_{\mathrm{stat}}$*, and* $\varepsilon_{\mathrm{opt}}$*, then the full-game Nash gap satisfies:*

$$\mathrm{Gap}_{\mathcal{G}}(\pi) \leq \tau + \varepsilon_{\mathrm{BR}} + \varepsilon_{\mathrm{plan}} + \varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{opt}}. \tag{11}$$

Furthermore, if $\mathcal{Q}$ *forms a* δ*-cover of a compact opponent class containing the deployed opponents, and both response loss and in-class deviation gain are controlled on that cover with Lipschitz constants* L_S *and* L_R*, the full-game Nash gap is bounded by* $\tau + \varepsilon_{\mathrm{BR}} + (L_S + L_R)\delta + \varepsilon_{\mathrm{plan}} + \varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{opt}}$.

Table 2: Operational decomposition of the unrestricted deviation guarantee.

Error Term	Operational Source	Control Mechanism
$\varepsilon_{\mathrm{BR}}$	Abstraction best-response loss	Joint Bellman optimization (Algorithm 1) or local certificate (10)
$(L_S + L_R)\delta$	Opponent-space approximation	Verified opponent cover
$\varepsilon_{\mathrm{plan}}$	Dynamic-programming approximation	Bellman residual control / discretization envelope β_M
$\varepsilon_{\mathrm{stat}}$	Empirical estimation error	Uniform concentration for empirical Bellman operators
$\varepsilon_{\mathrm{opt}}$	Numerical solver tolerance	Tree-elimination grid resolution M or branch-and-bound gap
τ	Restricted Nash gap of the meta-solver	PSRO / league meta-solver termination threshold

Table 3: Best-response loss under matched representation budgets. Lower is better.

Abstraction Criterion	Soccer ($K = 128$)	Connect Four ($K = 22{,}057$)	Connect Four Root Response Loss
Predictive Clustering	0.448	0.822	0.0714
All-Action Model Similarity	0.619	—	—
Value / Response-Q Clustering	0.280	0.457	0.0639
Equilibrium-Q Profile	0.421	—	—
Best-Response Abstraction (Ours)	**0.109**	$< 10^{-10}$	**0**

Table 2 summarizes this unified decomposition, separating representation error from opponent-cover, planning, statistical, and meta-solver terms.

7 Experiments

We evaluate global optimization on finite games, abstraction selection in Markov Soccer, and response compilation on Connect Four. Detailed environments, implementations, and additional experiments (including Tic-Tac-Toe PSRO and statistical coverage checks) appear in Appendix J.

Global Optimization on Finite Games. We evaluated Algorithm 1 on a suite of 76 finite discounted games with action-dependent transitions, comparing against exhaustive assignment search. The finite min–max objective values agreed with exhaustive grid enumeration in all 76 comparisons. On a representative three-state action-dependent game, the numerical optimization results give the bracket $[0.050510, 0.050522]$ of width 1.2×10^{-5}.

Markov Soccer under Opponent Population Shifts. We evaluate a 4×5 Markov Soccer benchmark (760 playable states, $\gamma = 0.9$, 38,000 transition branches). At a fixed budget of $K = 128$ abstract states, we benchmark a best-response abstraction against predictive clustering, all-action transition/reward similarity, value probes, and equilibrium-Q clustering across 32 opponent contexts. As reported in Table 3, the best-response abstraction achieves a best-response loss of 0.109, lower than predictive clustering (0.448) and value probes (0.280). In an expanding league pool of 48 opponents, adaptive opponent selection achieves a held-out best-response loss of 0.0825 using 240 states, whereas three random insertion orders average 0.1686 ± 0.0196 using 238.3 states on average (Figure 3(b)).

Scalable Compilation on 4×4 Connect Four. Connect Four features 161,029 reachable states and 304,574 transitions, cross-checked with independent coordinate-array and bitboard implementations. We evaluate whether a shared abstraction can compress the state space while preserving optimal counter-strategies against twelve benchmark opponents per player. Under a matched budget of $K = 22{,}057$ abstract states, all methods preserve the same zero-gap deployment. Best-response abstraction compiles the response policy set to numerical precision (best-response loss $< 10^{-10}$). In contrast, predictive clustering incurs a best-response loss of 0.8222 and centered response-Q clustering incurs 0.4573 (Table 3 and Figure 3(c)). The compressed representation consumes 31.8% of uncompressed deterministic policy storage while maintaining response fidelity to numerical precision.

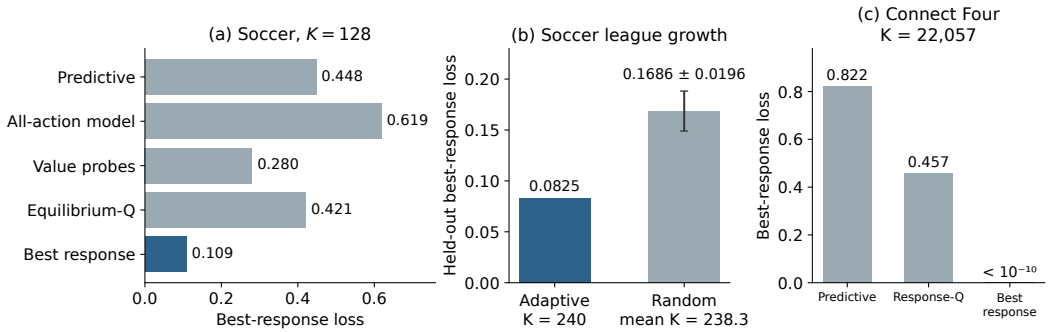

Figure 3: Best-response loss in the original games. (a) Markov Soccer at $K = 128$. (b) Held-out league evaluation: adaptive admission uses 240 states; three random orders use 238.3 on average, with reported loss 0.1686 ± 0.0196. (c) Connect Four at $K = 22{,}057$, comparing predictive clustering, centered response-Q clustering, and best-response refinement.

8 Discussion and Limitations

Our findings delineate the conditions under which strategy-complete state abstraction is most advantageous. When planning requires preservation of reward and transition behavior across many policies, model-based abstraction or bisimulation is appropriate. When the value evaluation class is fixed and static, value equivalence is well-suited. Strategy completeness is explicitly tailored for competitive multi-agent environments where the policy class must adapt to strategic opponent shifts under strict representation budgets.

The primary theoretical limitation of our global exact algorithm is its exponential scaling in the treewidth w of the primal graph induced by the joint Bellman factorization. When treewidth is large, exact global optimization becomes intractable, requiring recourse either to local spatial certification (Section 6.1) or mixed-integer outer relaxations (Appendix F). Establishing sample-complexity bounds under general non-linear function approximation remains an important open question.

9 Conclusion

In competitive multi-agent reinforcement learning, state abstraction must preserve the ability to respond effectively to changing opponents. We formulated *strategy completeness* as a quantitative objective for state abstraction in Markov games. Our exponential separation shows that preserving optimal counter-actions can require exponentially fewer abstract states than reconstructing the values of the same tasks, even under arbitrary action-dependent transitions. This identifies a distinct role for decision preservation when representation capacity is limited. We cast the joint synthesis of partitions and policy profiles as a factorized Bellman program. For fixed induced primal-graph treewidth, our algorithm computes certified ε-optimal abstractions in time polynomial in the explicit input size, H, and $1/\varepsilon$. Local certification under spatial contraction and a population error decomposition extend the framework to networked games, PSRO, and league training. Experiments on finite games, Markov Soccer, and 161,029-state Connect Four demonstrate lower benchmark response loss and response compilation to numerical precision under matched partition budgets. Future work includes certified abstraction in continuous state-action spaces with non-linear function approximation, and adaptive refinement as opponent populations grow. More broadly, the separation between decision preservation and value reconstruction motivates representations that allocate capacity to state distinctions needed by the current response tasks. This perspective connects compact policy compilation with scalable representation learning for competitive multi-agent systems.

AI USE STATEMENT

ChatGPT 5.6 Sol and 6 Pro were used to search for information and relevant literature, check mathematical arguments and proofs, and assist with manuscript consistency and presentation. Gemini 3.8 Flash was used for language polishing. The research topic, mathematical derivations, proofs, and experiments were developed and carried out by the human authors. Mathematical results were additionally checked using Wolfram, the Z3 solver, and Lean. The authors reviewed the AI-assisted work and take responsibility for the final text, claims, and results.

ETHICS STATEMENT

This work studies state abstraction in mathematical models and simulated games. The reported experiments involve no human participants or personal data. Competitive multi-agent systems can support applications with adversarial incentives; evaluating such applications requires attention to their deployment context.

REPRODUCIBILITY STATEMENT

Definitions, assumptions, and proofs are provided in Appendices C–I. Appendix J describes the experimental protocols, benchmark sizes, comparison budgets, and independent evaluation procedures. Algorithm 1 and Appendices E–F specify the optimization and certification constructions.

REFERENCES

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 2915–2923. PMLR, 2016.

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 10–19. PMLR, 2018.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 463–474. PMLR, 2020.

David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 434–443. PMLR, 2019.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 501–510. PMLR, 2018.

Antoine Bergerault, Volkan Cevher, and Negar Mehr. Matching multiple experts: On the exploitability of multi-agent imitation learning. In *International Conference on Learning Representations*, 2026.

Daniel Bienstock and Gonzalo Muñoz. LP approximations to mixed-integer polynomial optimization problems. *arXiv preprint arXiv:1501.00288*, 2015. Revised 2016.

Branislav Bošanský, Jiří Čermák, Karel Horák, and Michal Pěchouček. Computing maxmin strategies in extensive-form zero-sum games with imperfect recall. In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, pp. 63–74, 2017. doi: 10.5220/0006121200630074.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. MICo: Improved representations via sampling-based state similarity for markov decision processes. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jiří Čermák, Branislav Bošanský, and Viliam Lisý. An algorithm for constructing and solving imperfect recall abstractions of large extensive-form games. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 936–942, 2017. doi: 10.24963/ijcai.2017/130.

Sourav Chakraborty, Amit Kiran Rege, Claire Monteleoni, and Lijun Chen. A unified framework for locality in scalable MARL. In *Proceedings of the 8th Annual Learning for Dynamics and Control Conference*, volume 331, pp. 367–396. PMLR, 2026.

Qiwen Cui and Simon S. Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. In *Advances in Neural Information Processing Systems*, volume 35, 2022a. doi: 10.52202/068431-0853.

Qiwen Cui and Simon S. Du. When are offline two-player zero-sum markov games solvable? In *Advances in Neural Information Processing Systems*, volume 35, 2022b. doi: 10.52202/068431-1869.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5 (4):613–624, 1993. doi: 10.1162/neco.1993.5.4.613.

Thomas Dean, Robert Givan, and Sonia Leach. Model reduction techniques for computing approximately optimal solutions for markov decision processes. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp. 124–131, 1997.

Thomas L. Dean and Robert Givan. Model minimization in markov decision processes. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 106–111, 1997.

Rina Dechter. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, 113(1–2):41–85, 1999. doi: 10.1016/S0004-3702(99)00059-4.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1665–1674. PMLR, 2019.

Yuri Faenza, Gonzalo Muñoz, and Sebastian Pokutta. New limits of treewidth-based tractability in optimization. *arXiv preprint arXiv:1807.02551*, 2018. Revised 2019.

Amir Massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1486–1494. PMLR, 2017.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pp. 950–951, 2004.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484X.

A. M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I*, 28:89–93, 1964. doi: 10.32917/hmj/1206139508.

Vincent François-Lavet, Yoshua Bengio, Doina Precup, and Joelle Pineau. Combined reinforcement learning via abstract representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3582–3589, 2019. doi: 10.1609/aaai.v33i01.33013582.

Till Freihaut, Luca Viano, Volkan Cevher, Matthieu Geist, and Giorgia Ramponi. Learning equilibria from data: Provably efficient multi-agent imitation learning. In *Advances in Neural Information Processing Systems*, 2025.

Till Freihaut, Luca Viano, Emanuele Nevali, Volkan Cevher, Matthieu Geist, and Giorgia Ramponi. Rate optimal learning of equilibria from data. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300, pp. 3430–3438. PMLR, 2026.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2170–2179. PMLR, 2019.

Andrew Gilpin and Tuomas Sandholm. A competitive texas hold'em poker player via automated abstraction and real-time equilibrium computation. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pp. 1007–1013, 2006.

Andrew Gilpin and Tuomas Sandholm. Better automated abstraction techniques for imperfect information games, with application to texas hold'em poker. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, 2007a.

Andrew Gilpin and Tuomas Sandholm. Lossless abstraction of imperfect information games. *Journal of the ACM*, 54(5), 2007b. doi: 10.1145/1284320.1284324.

Andrew Gilpin, Tuomas Sandholm, and Troels Bjerre Sørensen. Potential-aware automated abstraction of sequential games, and holistic equilibrium analysis of texas hold'em poker. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pp. 50–57, 2007.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Marcus Hutter. Extreme state aggregation beyond MDPs. In *Algorithmic Learning Theory*, pp. 185–199, 2014. doi: 10.1007/978-3-319-11662-4_14.

Hiroki Ishibashi, Kenshi Abe, and Atsushi Iwasaki. Approximate state abstraction for markov games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17555–17563, 2025. doi: 10.1609/aaai.v39i17.33930.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 179–188. PMLR, 2015.

Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Michael Kearns, Michael L. Littman, and Satinder Singh. Graphical models for game theory. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 253–260, 2001.

Christian Kroer and Tuomas Sandholm. Extensive-form game abstraction with bounds. In *Proceedings of the 15th ACM Conference on Economics and Computation*, pp. 621–638, 2014. doi: 10.1145/2600057.2602905.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4190–4203, 2017.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Thomas Ewalds, Ryan Faulkner, Janos Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, 2006.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I—convex underestimating problems. *Mathematical Programming*, 10:147–175, 1976. doi: 10.1007/BF01580665.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6961–6971. PMLR, 2020.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In *Algorithmic Learning Theory*, pp. 140–154, 2014. doi: 10.1007/978-3-319-11662-4_11.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628, 2022. doi: 10.1287/opre.2021.2226.

Balaraman Ravindran and Andrew G. Barto. SMDP homomorphisms: An algebraic approach to abstraction in semi-markov decision processes. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 1011–1016, 2003.

Zhaolin Ren, Runyu Zhang, Bo Dai, and Na Li. Scalable spectral representations for multiagent reinforcement learning in network MDPs. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 550–558. PMLR, 2025.

Naci Saldi, Serdar Y"uksel, and Tamás Linder. On the asymptotic optimality of finite approximations to markov decision processes with borel spaces. *Mathematics of Operations Research*, 42(4):945–978, 2017. doi: 10.1287/moor.2016.0832.

Tuomas Sandholm and Satinder Singh. Lossy stochastic game abstraction with bounds. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 880–897, 2012. doi: 10.1145/2229012.2229079.

Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39:1095–1100, 1953. doi: 10.1073/pnas.39.10.1095.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pp. 512–519, 2004.

Elise van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. MDP homomorphic networks: Group symmetries in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Luca Viano, Till Freihaut, Emanuele Nevali, Volkan Cevher, Matthieu Geist, and Giorgia Ramponi. Multi-agent imitation learning with function approximation: Linear markov games and beyond. *arXiv preprint arXiv:2602.22810*, 2026.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019. doi: 10.1038/s41586-019-1724-z.

Wenhao Zhan, Scott Fujimoto, Zheqing Zhu, Jason D. Lee, Daniel R. Jiang, and Yonathan Efroni. Exploiting structure in offline multi-agent RL: The benefits of low interaction rank. In *International Conference on Learning Representations*, 2025.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Yuheng Zhang, Yu Bai, and Nan Jiang. Offline learning in markov games with general function approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 40804–40829. PMLR, 2023.

Zhaoyi Zhou, Zaiwei Chen, Yiheng Lin, and Adam Wierman. Convergence rates for localized actor-critic in networked markov potential games. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 2563–2573. PMLR, 2023.

APPENDIX OVERVIEW

Appendices A–B provide notation and related work; Appendices C–F give the joint formulation, separation results, and optimization proofs; Appendices G–I develop statistical, population, and local guarantees. Appendix J describes the experimental protocols and results.

APPENDIX CONTENTS

A. Notation and terminology.

B. Detailed related work: model/bisimulation, value equivalence, game abstraction, strategy completeness, and population-based MARL.

C. Best-response loss, strategy completeness, equilibrium-transfer error, and the exact joint Bellman formulation.

D. Exponential separation between decision and value abstraction, perturbation robustness, response conflicts, and analytical counterexamples.

E. Bounded-treewidth joint optimization: canonical discretization, outward rounding, finite decoding, and min–max variable elimination.

F. Mixed-integer outer relaxation and structural optimization for high-treewidth games.

G. Statistical guarantees and offline sample complexity: uniform empirical Bellman concentration, population optimality intervals, and query lower bounds.

H. Population-based learning and opponent space representation: episode mixtures, public-history Markov states, and metric opponent coverage.

I. Local certification of best responses in networked games: truncated local MDPs, boundary influence error, spatial decay, and block-invariant decompositions.

J. Experimental protocols, benchmark specifications, and extended empirical results for Soccer, Connect Four, Tic-Tac-Toe PSRO, and exact finite games.

A NOTATION AND TERMINOLOGY

Throughout this paper, we consider a finite discounted n-player Markov game $\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{r_i\}_{i\in\mathcal{N}}, \gamma)$ and a designated finite benchmark opponent set $\mathcal{Q} \subset \Pi_{-i}$, unless explicitly stated otherwise. We write D_K and J_K as shorthands for $D_K(\mathcal{G}, \mathcal{Q})$ and $J_K(\mathcal{G}, \mathcal{Q})$ whenever the environment and task arguments are clear from context.

Best-Response Loss and Strategy Completeness. For a specific response task induced by stationary opponent strategy $q \in \Pi_{-i}$, the focal player i faces an effective single-agent MDP. While an unrestricted deviation may condition on the full original state history, let W_i^q denote the unrestricted optimal value vector satisfying $W_i^q = \mathcal{T}_i^{*,q} W_i^q$, and let $V_i^{\sigma,q}$ denote the evaluated return under a stationary ϕ-abstract policy $\sigma \in \Pi_{\phi,i}$. The best-response loss is defined as:

$$S_q(\phi) = \inf_{\sigma\in\Pi_{\phi,i}} \|W_i^q - V_i^{\sigma,q}\|_\infty.$$

The restricted policy class $\Pi_{\phi,i}$ is strategy-complete for task q if and only if $S_q(\phi) = 0$. Strategy completeness thus constitutes a qualitative preservation property, whereas $S_q(\phi)$ provides its quantitative defect used for approximate state abstraction under finite representation budgets.

Opponent-Conditioned Abstraction Loss and Equilibrium Transfer. For a designated benchmark opponent pool $\mathcal{Q} \subset \Pi_{-i}$, the aggregate abstraction loss is:

$$S_{\mathcal{Q}}(\phi) = \max_{q\in\mathcal{Q}} S_q(\phi), \qquad D_K(\mathcal{G}, \mathcal{Q}) = \min_{|\phi(\mathcal{S})|\le K} S_{\mathcal{Q}}(\phi).$$

The equilibrium-transfer error of partition ϕ is:

$$E_\phi(\mathcal{G}) = \inf_{\pi\in\Pi_\phi} \mathrm{Gap}_{\mathcal{G}}(\pi).$$

Table 4: Summary of mathematical notation, operations, and technical terminology.

Term	Symbol	Formal Definition / Meaning
State partition	$\phi : \mathcal{S} \to [K]$	Mapping partitioning the original state space into at most K cells.
ϕ-Abstract policy class	Π_ϕ	Subclass of stationary Markov policies constant across states sharing partition cells: $\phi(s) = \phi(s') \implies \pi(\cdot \mid s) = \pi(\cdot \mid s')$.
Best-response loss	$S_q(\phi)$	Uniform value loss incurred by restricting player i's best response against opponent profile q to $\Pi_{\phi,i}$: $\inf_{\sigma \in \Pi_{\phi,i}} \|W_i^q - V_i^{\sigma,q}\|_\infty$.
Strategy completeness	$S_q(\phi) = 0$	Structural property that the restricted policy class $\Pi_{\phi,i}$ contains an exact unrestricted best response against opponent q.
Opponent-conditioned abstraction loss	$D_K(\mathcal{G}, \mathcal{Q})$	Minimum worst-case best-response loss over all state partitions of cardinality at most K: $\min_{\lvert\phi(\mathcal{S})\rvert \le K} \max_{q \in \mathcal{Q}} S_q(\phi)$.
Nash gap	$\mathrm{Gap}_{\mathcal{G}}(\pi)$	Worst-player, worst-state unilateral deviation gain in the original game: $\max_{i \in \mathcal{N}} \|W_i^{\pi_{-i}} - V_i^{\pi}\|_\infty$.
Equilibrium-transfer error	$E_\phi(\mathcal{G})$	Infimum of the Nash gap achievable by any deployment profile supported on partition ϕ: $\inf_{\pi \in \Pi_\phi} \mathrm{Gap}_{\mathcal{G}}(\pi)$.
Joint abstraction objective	$J_K(\mathcal{G}, \mathcal{Q})$	Simultaneous minimax optimum over one shared state partition ϕ of deployment error and declared task response losses: $\min_{\lvert\phi(\mathcal{S})\rvert \le K} \max\{E_\phi(\mathcal{G}), S_{\mathcal{Q}}(\phi)\}$.
Opponent metric coverage	δ-cover of $\mathcal{Q}_{\mathrm{all}}$	Approximation of a larger compact opponent class by a finite benchmark pool under statewise total variation distance.
Local best-response certificate	$H[\alpha_f + \beta_{f,\sigma} + 2e_B(f)]$	Certified upper bound on best-response loss obtained from a truncated local MDP and spatial boundary leakage bounds.
Joint Bellman factor graph	$\mathcal{G}_{\mathrm{B}}$	Factor graph representation capturing the coupling between partition assignments, mixed action probabilities, and continuation values.
Treewidth	w	Treewidth of the primal graph induced by the joint Bellman factorization; factors connect all variables in their scopes.

Here, $\mathrm{Gap}_{\mathcal{G}}(\pi)$ denotes the Nash gap of profile π evaluated on the ground game, while $E_\phi(\mathcal{G})$ isolates the best deployment profile representable within Π_ϕ. The joint abstraction objective simultaneously optimizes both requirements under a common partition:

$$J_K(\mathcal{G}, \mathcal{Q}) = \min_{|\phi(\mathcal{S})| \le K} \max\left\{E_\phi(\mathcal{G}),\ S_{\mathcal{Q}}(\phi)\right\}.$$

Transition and Model Dynamics. An action-independent or exogenous transition in an induced response MDP satisfies $P(\cdot \mid s, a_i) = P(\cdot \mid s)$ for all $a_i \in \mathcal{A}_i$. A near-exogenous transition is uniformly close in total variation distance to an action-independent transition kernel. While these special cases simplify computational analysis in linear-programming regimes, Theorem 4.1, Theorem 5.1, and the mixed-integer formulation of Appendix F establish guarantees for arbitrary action-dependent transition dynamics without structural simplifications.

Valuation Scales and Policy Semantics. All principal error metrics and objective values are measured in worst-state discounted return units bounded by the effective horizon $H := (1-\gamma)^{-1}$. Guarantees under specific initial distributions $\mu \in \Delta(\mathcal{S})$ are designated as such and are strictly differentiated from uniform sup-norm certificates. The representation budget K denotes the discrete cardinality bound on the abstract state space (each label is also called a code) $|\phi(\mathcal{S})| \le K$; it is conceptually distinct from the latent continuous embedding dimension, neural network parameter capacity, or total memory footprint. In our empirical evaluations, the state-to-partition lookup index and all task-specific response policy heads are fully accounted for in the actor memory footprint.

A benchmark opponent profile q specifies an explicit response task in $\mathcal{Q}$. In PSRO or league training, admitting a novel strategy profile augments the task pool $\mathcal{Q}$. When several focal players are tracked simultaneously, one may equivalently index response tasks by $h = (i, q)$. A stationary Markov mixture randomizes actions statewise conditional solely on the current state. In contrast, an episode-level mixture samples an opponent policy from a population distribution once at episode inception; optimal responses against episode mixtures require conditioning on the public interaction history, or its Bayesian belief state representation, as formalized in Appendix H.

B Detailed Related Work

State Abstraction in Single-Agent MDPs. Classical state aggregation and stochastic bisimulation metrics merge environment states whose local reward distributions and action-conditioned transition probabilities coincide after quotient reduction (Dean & Givan, 1997; Dean et al., 1997; Givan et al., 2003; Ravindran & Barto, 2003). Approximate abstractions replace strict equality with value- or model-error tolerances (Ferns et al., 2004; 2011; Abel et al., 2016; Hutter, 2014); abstraction selection and finite approximations study how much aggregation can be tolerated while retaining near-optimal control (Jiang et al., 2015; Ortner et al., 2014; Saldi et al., 2017; Abel et al., 2018). Deep latent representation learning replaces explicit combinatorial partitions with learned encoders and predictive or behavioral objectives (Gelada et al., 2019; François-Lavet et al., 2019; Zhang et al., 2021; Castro et al., 2021); related rich-observation and invariance methods learn latent state structure from high-dimensional observations (Du et al., 2019; Misra et al., 2020; Agarwal et al., 2020; van der Pol et al., 2020). These classical criteria are environment-facing: by preserving transition and reward information, they support planning for compatible tasks and can retain distinctions unnecessary for a particular opponent family.

Decision-Aware Representations and Value Equivalence. Predictive-state and successor representations summarize future observations or occupancies rather than reconstructing the raw state (Littman et al., 2001; Singh et al., 2004; Dayan, 1993); successor features extend this idea across reward functions (Barreto et al., 2017; 2018). Value-aware model learning and value-targeted regression retain only the transition components needed by selected value computations (Farahmand et al., 2017; Ayoub et al., 2020). The value-equivalence principle formalizes this paradigm relative to pre-specified policy classes and value-function probe families (Grimm et al., 2020). When the probe family is narrow, value-equivalent models achieve aggressive compression compared to full transition models (Abel et al., 2018). However, best-response preservation introduces a different preservation target. Rather than ensuring that the numerical evaluations of a static probe class match the ground truth, strategy completeness demands that the compressed policy class contains the optimal decision boundaries required to best-respond against an opponent family. As proved in Theorem 4.1, tracking the numerical values of a task family can require exponentially more abstract states than identifying its optimal counter-actions.

Game Abstraction and Equilibrium Transfer. Abstraction techniques in extensive-form and stochastic games have historically focused on computing reduced games or strategically similar information states (Gilpin & Sandholm, 2006; 2007a;b; Gilpin et al., 2007). Regret and solution-quality bounds connect the reduced game back to the original one (Kroer & Sandholm, 2014; Sandholm & Singh, 2012; Ishibashi et al., 2025). Imperfect-recall work provides complementary algorithmic perspectives on compact game representations and refinement (Čermák et al., 2017). In Markov games, recent abstraction frameworks bound the ground-game exploitability of lifted abstract equilibria based on state similarity metrics (Ishibashi et al., 2025). These approaches focus primarily on deployment fidelity, corresponding to the equilibrium-transfer error $E_\phi(\mathcal{G})$ in our framework. Preserving an equilibrium strategy profile does not guarantee that the compressed representation retains the strategic flexibility to formulate counter-strategies against novel, out-of-equilibrium opponents encountered during deployment or training. Strategy-complete abstraction addresses this limitation by jointly optimizing deployment fidelity alongside multi-task best-response capabilities $S_{\mathcal{Q}}(\phi)$.

Strategy Completeness and Unilateral Deviation Coverage in Offline MARL. In offline multi-agent learning and batch policy optimization, equilibrium-path coverage alone is insufficient to prevent

out-of-distribution exploitation (Cui & Du, 2022b;a). Function-approximation and low-interaction-rank analyses provide complementary structural routes to learning under restricted data (Zhang et al., 2023; Zhan et al., 2025). Recent equilibrium and imitation-learning results sharpen the role of deviation coverage and feature structure (Freihaut et al., 2025; 2026; Bergerault et al., 2026; Viano et al., 2026). While prior literature treats strategy completeness as a structural assumption imposed on an abstract policy class, our work operationalizes completeness as a constructive, optimization-driven objective for state abstraction under explicit partition cardinality constraints.

Population-Based Training and Opponent Diversity. Policy-Space Response Oracles (PSRO) and league training expand an empirical population through response oracles (Lanctot et al., 2017; Balduzzi et al., 2019; Vinyals et al., 2019). General game tooling provides a complementary implementation framework (Lanctot et al., 2019). In these regimes, the downstream optimization task continually shifts as the meta-solver identifies new strategically relevant opponents. Existing methods compress policies via generic neural network parameterizations, which provide no formal guarantees against representation-induced exploitability. Strategy-complete state abstraction provides a formal representational interface for PSRO and league training: by monitoring the best-response loss $S_q(\phi)$, the abstraction module determines precisely when a newly admitted opponent requires refining the state partition, isolating representation error from meta-solver population regret.

Graph-Theoretic Structure and Networked MARL. Networked reinforcement learning exploits spatial decay and local factorizations (Qu et al., 2022; Zhou et al., 2023; Ren et al., 2025; Chakraborty et al., 2026). Our global optimization framework leverages graph structure at a different level: the factor graph of the joint abstraction-and-policy optimization program $\mathcal{G}_{\mathrm{B}}$. Exact variable elimination and bounded-treewidth dynamic programming provide algorithms for graphical games and structured optimization (Kearns et al., 2001; Dechter, 1999; Bienstock & Muñoz, 2015; Faenza et al., 2018). Mixed-integer and refinement methods have also been used for compact controllers and imperfect-recall game abstractions (Bošanský et al., 2017; Čermák et al., 2017; McCormick, 1976). Theorem 5.1 establishes a provably convergent algorithm for joint state abstraction and policy optimization in Markov games with bounded induced primal-graph treewidth.

B.1 Comparison of Preservation Objectives

State abstraction criteria differ in the mathematical invariant that is preserved under state aggregation. Two states that are distinct under one criterion may be merged safely under another. Table 5 summarizes these preservation paradigms under a unified formal framing.

Table 5: Comparison of state-abstraction paradigms in Markov games. Strategy completeness is the zero-loss property quantified by the best-response loss $S_q(\phi)$.

Abstraction Paradigm	Mathematical Invariant / Exact Condition	Approximate Guarantee	Operational Implication for Changing Opponents
Exact Model Aggregation	Rewards and probabilities of each abstract successor cell agree across merged states for every action	Simulation bounds and near-optimal planning via model approximation	Reusable across all downstream tasks; retains state distinctions that never alter competitive responses.
Bisimulation Metric	Recursive metric distance vanishes: $d_{\sim}(s, s') = 0$	Value-error bounds under the metric's policy and normalization assumptions	Preserves universal behavioral similarity; may retain distinctions unnecessary for a specific opponent pool.
Value Equivalence	Commuting Bellman evaluation on probe class: $\mathcal{T}^{\pi} V = \mathcal{T}^{\pi}_{\mathrm{model}} V$ for $\pi \in \Pi_{\mathrm{probe}}, V \in \mathcal{V}_{\mathrm{probe}}$	Policy evaluation error bounded by probe-class learning loss	Task-directed; retaining numerical values for broad probe classes preserves payoff shifts that do not impact action optimality.
Equilibrium Abstraction	Exact equilibrium lifting: $\mathrm{Gap}_{\mathcal{G}}(\pi) = 0$ for abstract equilibrium $\pi \in \Pi_{\phi}$	Bounded original-game exploitability of lifted reduced-game equilibrium	Directly protects the deployed profile; may omit responses required against other opponents.
Strategy Completeness (Ours)	Best-response operator fixed point: $\mathcal{T}_i^{*,q} W_i^q = \mathcal{T}_i^{\sigma_q^*,q} W_i^q$ with $\sigma_q^* \in \Pi_{\phi,i}$	Unrestricted deviation gain bounded by restricted regret plus $S_q(\phi)$	Representation capacity scales directly with the strategic diversity of the opponent pool $\mathcal{Q}$.

B.2 Relation to Exact and Approximate State Abstraction

Classical state aggregation and MDP homomorphisms merge states that exhibit identical transition and reward behavior under all actions (Givan et al., 2003; Ravindran & Barto, 2003; Li et al., 2006).

Approximate variants replace strict equivalence with bounded metric discrepancy, bounding the value loss incurred when lifting an abstract policy back to the original environment (Abel et al., 2016; 2018; Ferns et al., 2004; 2011).

For any fixed opponent profile q, a model-homomorphic abstraction of the induced single-agent response MDP preserves the Bellman optimality fixed point, directly implying $S_q(\phi) = 0$. However, the converse implication is false. As established in Lemma D.1, achieving strategy completeness $S_q(\phi) = 0$ requires only that all states within a partition cell share at least one common optimal counter-action. The underlying transition probabilities, reward magnitudes, and continuation values may vary substantially within the cell. Theorem 4.1 proves that this distinction can yield an exponential gap in representation complexity under identical benchmark tasks.

B.3 Relation to Value Equivalence

The value-equivalence principle relaxes full environment modeling by requiring only that Bellman evaluations are preserved for a designated class of policies and value functions (Grimm et al., 2020). While value equivalence is already task-directed, Theorem 4.1 isolates the information gap between value reconstruction and decision preservation.

Specifically, consider a fixed opponent family $\mathcal{Q}$. Strategy completeness requires the abstract state partition ϕ to preserve only the optimal action choices against $\mathcal{Q}$. In contrast, a value-equivalent representation that reconstructs the optimal value functions W^q must track all payoff-shifting environmental variations, even when those variations do not alter the argmax of the Bellman optimality operator. Theorem 4.1 shows that under identical dynamics, achieving tolerance ε_0 requires $2^{|T|}$ abstract states for decision preservation versus 2^{m+n_0} abstract states for numerical value reconstruction.

B.4 Relation to Game Abstraction and Equilibrium Transfer

Stochastic and extensive-form game abstractions aggregate game states to solve a smaller abstract game whose computed equilibrium lifts back to the ground game with bounded regret (Sandholm & Singh, 2012; Kroer & Sandholm, 2014; Ishibashi et al., 2025). For a candidate partition ϕ, the equilibrium-transfer error $E_\phi(\mathcal{G})$ evaluates precisely this deployment-side fidelity.

Equilibrium transfer controls deviations against the lifted profile, rather than best responses against every benchmark opponent. A compressed representation can achieve zero equilibrium-transfer error $E_\phi(\mathcal{G}) = 0$ while incurring best-response loss $\Omega(H)$ when facing a fixed non-equilibrium opponent (as shown in Counterexample 2, Appendix D.4). Strategy-complete abstraction addresses this vulnerability by explicitly formulating the dual objective: the joint program $J_K(\mathcal{G}, \mathcal{Q})$ optimizes a single shared partition ϕ that simultaneously minimizes deployment Nash gap $E_\phi(\mathcal{G})$ and maintains counter-strategy flexibility $S_\mathcal{Q}(\phi)$ against the opponent benchmark pool.

B.5 Relation to Strategy Completeness and Offline Deviation Coverage

Theoretical analyses of offline Markov games establish that data coverage over the equilibrium path is insufficient to guarantee low Nash gap against arbitrary deviations (Cui & Du, 2022b; Zhang et al., 2023). To address this, existing literature imposes strategy completeness as an axiomatic assumption on the hypothesis policy class. The partition mapping ϕ operationalizes this concept constructively: it defines the restricted policy space Π_ϕ, while $S_q(\phi)$ measures its quantitative defect.

Strategy completeness is an intrinsic representational property of the policy class, distinct from statistical data coverage. An abstract representation can be strategy-complete in the true game even when the available offline dataset lacks the coverage required to identify the model parameters. Appendix G formally decouples statistical model-estimation error from abstraction incompleteness error.

B.6 Computational Paradigms and Complexity Trade-offs

When the state aggregation criterion is fixed, tabular single-agent abstractions can often be computed via polynomial-time partition refinement algorithms. In multi-agent Markov games, optimizing

the joint objective $J_K(\mathcal{G}, \mathcal{Q})$ is non-convex due to the bilinear coupling between discrete partition variables $\phi(s) \in [K]$, continuous mixed action probabilities, and continuation values.

Our framework establishes three complementary computational regimes for optimizing $J_K(\mathcal{G}, \mathcal{Q})$:

1. **Bounded-Treewidth Dynamic Programming (Algorithm 1, Appendix E):** When the induced primal graph of $\mathcal{G}_{\mathrm{B}}$ has bounded treewidth w, min–max bucket elimination on the discretized program computes a certified ε-optimal abstraction in polynomial time $\mathrm{poly}(\mathcal{I}, \log M) \max\{K, M+1\}^{w+1}$.
2. **Mixed-Integer Outer Relaxation (Appendix F):** For general games where treewidth is large, the non-convex program is cast as a mixed-integer nonlinear program with McCormick linear envelopes, solved via branch-and-bound to yield globally valid lower and upper bounds.
3. **Local Certification (Appendix I):** For large-scale factored networked systems, solving a truncated local MDP on a sub-network yields a certified best-response upper bound without requiring global factor graph elimination.

C Best-Response Loss and Exact Joint Bellman Formulation

This appendix provides the formal foundational analysis for the best-response loss, the strategy-completeness property, and the exactness of the joint Bellman optimization formulation introduced in Section 3.

C.1 Notation and Induced Environment Operators

Let $\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{r_i\}_{i\in\mathcal{N}}, \gamma)$ be a finite discounted n-player Markov game with state space cardinality $|\mathcal{S}|$, joint action space $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}_i$, and effective horizon $H = (1-\gamma)^{-1}$. Let $\mathcal{Q} \subset \Pi_{-i}$ be a finite benchmark set of stationary opponent profiles. Each opponent profile $q \in \mathcal{Q}$ specifies a response task for player i, wherein player i seeks to optimize against the fixed environment induced by q.

For a fixed opponent profile $q \in \Pi_{-i}$, the induced stage reward and transition kernel for player i are denoted by $r_i^q(s, a_i) := \mathbb{E}_{\mathbf{a}_{-i}\sim q(\cdot|s)}[r_i(s, a_i, \mathbf{a}_{-i})]$ and $P_i^q(s' \mid s, a_i) := \mathbb{E}_{\mathbf{a}_{-i}\sim q(\cdot|s)}[P(s' \mid s, a_i, \mathbf{a}_{-i})]$. For any bounded value vector $v \in \mathbb{R}^{|\mathcal{S}|}$, we define the corresponding Bellman operators:

$$(\mathcal{T}_i^{a_i,q} v)(s) := r_i^q(s, a_i) + \gamma \sum_{s'\in\mathcal{S}} P_i^q(s' \mid s, a_i) v(s'), \tag{12}$$

$$(\mathcal{T}_i^{*,q} v)(s) := \max_{a_i\in\mathcal{A}_i} (\mathcal{T}_i^{a_i,q} v)(s), \tag{13}$$

$$(\mathcal{T}_i^{\sigma,q} v)(s) := \sum_{a_i\in\mathcal{A}_i} \sigma(a_i \mid s)(\mathcal{T}_i^{a_i,q} v)(s), \tag{14}$$

where $\sigma \in \Pi_i$ is a stationary Markov response policy.

Lemma C.1 (Contraction and Best-Response Characterization)**.** *For any fixed opponent profile $q \in \Pi_{-i}$ and any stationary response policy $\sigma \in \Pi_i$:*

1. *The policy evaluation operator $\mathcal{T}_i^{\sigma,q}$ is a γ-contraction in the supremum norm with unique fixed point $V_i^{\sigma,q}$.*
2. *The Bellman optimality operator $\mathcal{T}_i^{*,q}$ is a γ-contraction in the supremum norm with unique fixed point W_i^q.*
3. *There exists a deterministic stationary policy $\sigma^* \in \Pi_i$ that attains W_i^q simultaneously across all states $s \in \mathcal{S}$, and W_i^q dominates the expected return of any history-dependent policy.*
4. *If $v = \mathcal{T}_i^{\sigma,q} v$ and $u \geq \mathcal{T}_i^{*,q} u$ coordinatewise, then $u \geq W_i^q \geq v$.*
5. *The value functions $V_i^{\sigma,q}$ and W_i^q are continuous with respect to policy parameters, and the minima defining $S_q(\phi)$, $D_K(\mathcal{G}, \mathcal{Q})$, and $J_K(\mathcal{G}, \mathcal{Q})$ are achieved.*

Proof. We establish each property sequentially.

Step 1: Contraction of fixed-action operators. For any $v, w \in \mathbb{R}^{|\mathcal{S}|}$ and $a_i \in \mathcal{A}_i$,

$$|(\mathcal{T}_i^{a_i,q}v)(s) - (\mathcal{T}_i^{a_i,q}w)(s)| = \gamma \left| \sum_{s' \in \mathcal{S}} P_i^q(s' \mid s, a_i)(v(s') - w(s')) \right| \leq \gamma \sum_{s' \in \mathcal{S}} P_i^q(s' \mid s, a_i)\|v - w\|_\infty = \gamma \|v - w\|_\infty.$$

Taking the supremum over $s \in \mathcal{S}$ yields $\|\mathcal{T}_i^{a_i,q}v - \mathcal{T}_i^{a_i,q}w\|_\infty \leq \gamma\|v - w\|_\infty$.

Step 2: Contraction of policy evaluation. For any stationary policy $\sigma \in \Pi_i$, using $\sum_{a_i \in \mathcal{A}_i} \sigma(a_i \mid s) = 1$ and $\sigma(a_i \mid s) \geq 0$,

$$|(\mathcal{T}_i^{\sigma,q}v)(s) - (\mathcal{T}_i^{\sigma,q}w)(s)| \leq \sum_{a_i \in \mathcal{A}_i} \sigma(a_i \mid s)|(\mathcal{T}_i^{a_i,q}v)(s) - (\mathcal{T}_i^{a_i,q}w)(s)| \leq \gamma\|v - w\|_\infty.$$

Thus, $\mathcal{T}_i^{\sigma,q}$ is a γ-contraction in supremum norm.

Step 3: Contraction of optimality operator. For any real-valued families $\{x_a\}_{a \in \mathcal{A}_i}$ and $\{y_a\}_{a \in \mathcal{A}_i}$, the identity $|\max_a x_a - \max_a y_a| \leq \max_a |x_a - y_a|$ holds. Setting $x_a = (\mathcal{T}_i^{a,q}v)(s)$ and $y_a = (\mathcal{T}_i^{a,q}w)(s)$ yields

$$|(\mathcal{T}_i^{*,q}v)(s) - (\mathcal{T}_i^{*,q}w)(s)| \leq \max_{a_i \in \mathcal{A}_i} |(\mathcal{T}_i^{a_i,q}v)(s) - (\mathcal{T}_i^{a_i,q}w)(s)| \leq \gamma\|v - w\|_\infty,$$

which proves $\|\mathcal{T}_i^{*,q}v - \mathcal{T}_i^{*,q}w\|_\infty \leq \gamma\|v - w\|_\infty$.

Step 4: Existence, uniqueness, and invariant bounds. By the Banach fixed-point theorem, both $\mathcal{T}_i^{\sigma,q}$ and $\mathcal{T}_i^{*,q}$ possess unique fixed points, denoted by $V_i^{\sigma,q}$ and W_i^q respectively. Given $r_i \in [0, 1]$, if $0 \leq v \leq H\mathbf{1}$, then $0 \leq \mathcal{T}_i^{a_i,q}v \leq 1 + \gamma H = H$. Thus, the compact hypercube $[0, H]^{|\mathcal{S}|}$ is invariant under both operators, ensuring that all fixed points lie within this domain.

Step 5: Domination over history-dependent deviations. Let π^{hist} be an arbitrary history-dependent policy choosing action A_t at step t. For any bounded vector v, with $\mathcal{F}_t$ including the selected action, conditional expectation satisfies

$$\mathbb{E}\left[R_t + \gamma v(S_{t+1}) - v(S_t) \mid \mathcal{F}_t\right] = (\mathcal{T}_i^{A_t,q}v)(S_t) - v(S_t).$$

Multiplying by γ^t, summing over $t = 0, \ldots, T-1$, and taking expectations from state s yields the telescoping identity:

$$\mathbb{E}_s\left[\sum_{t=0}^{T-1} \gamma^t \Big((\mathcal{T}_i^{A_t,q}v)(S_t) - v(S_t)\Big)\right] = \mathbb{E}_s\left[\sum_{t=0}^{T-1} \gamma^t R_t\right] - v(s) + \gamma^T \mathbb{E}_s[v(S_T)].$$

Setting $v = W_i^q$, we have $(\mathcal{T}_i^{a_i,q}W_i^q)(s) \leq (\mathcal{T}_i^{*,q}W_i^q)(s) = W_i^q(s)$ for all $a_i \in \mathcal{A}_i$, implying that the left-hand side is nonpositive. Thus:

$$\mathbb{E}_s\left[\sum_{t=0}^{T-1} \gamma^t R_t\right] \leq W_i^q(s) - \gamma^T \mathbb{E}_s[W_i^q(S_T)] \leq W_i^q(s).$$

Taking $T \to \infty$ confirms that the expected discounted return of any history-dependent policy cannot exceed $W_i^q(s)$.

Step 6: Existence of an optimal deterministic policy. Define $a^*(s) \in \arg\max_{a_i \in \mathcal{A}_i}(\mathcal{T}_i^{a_i,q}W_i^q)(s)$ for each state $s \in \mathcal{S}$. Then $\mathcal{T}_i^{a^*,q}W_i^q = \mathcal{T}_i^{*,q}W_i^q = W_i^q$. Uniqueness of the fixed point of $\mathcal{T}_i^{a^*,q}$ implies $V_i^{a^*,q} = W_i^q$ uniformly across all states.

Step 7: Supersolution monotonicity. If $u \geq \mathcal{T}_i^{*,q}u$, monotonicity of $\mathcal{T}_i^{*,q}$ implies $u \geq \mathcal{T}_i^{*,q}u \geq (\mathcal{T}_i^{*,q})^2u \geq \cdots \geq \lim_{k\to\infty}(\mathcal{T}_i^{*,q})^k u = W_i^q$. Conversely, if $v = \mathcal{T}_i^{\sigma,q}v$, uniqueness ensures $v = V_i^{\sigma,q} \leq W_i^q$.

Step 8: Continuity. The value of a stationary policy satisfies $V_i^{\sigma,q} = (I - \gamma P_i^{\sigma,q})^{-1} r_i^{\sigma,q} = \sum_{t=0}^{\infty} \gamma^t (P_i^{\sigma,q})^t r_i^{\sigma,q}$. The resolvent operator exists because the spectral radius satisfies $\rho(\gamma P_i^{\sigma,q}) \leq \gamma < 1$, and its induced operator norm is bounded by H. Since finite-state transition matrices and

reward vectors are continuous in policy parameters, $V_i^{\sigma,q}$ is continuous in σ. The fixed point W_i^q is likewise continuous with respect to model parameters by contraction perturbation bounds.

Step 9: Attainment of optima. For any fixed partition ϕ, the space of ϕ-abstract policies Π_ϕ is a compact product of simplexes. The continuity established in Step 8 guarantees that the infimum over $\sigma \in \Pi_{\phi,i}$ is attained. Because the collection of all partitions $\phi : \mathcal{S} \to [K]$ is finite, the outer minimum is achieved. □

C.2 Existence of Stationary Equilibria in Discounted Markov Games

Finite discounted general-sum stochastic games admit stationary Markov equilibria (Fink, 1964), establishing the foundational dynamic programming basis originally developed for zero-sum games by Shapley (1953). For any stationary opponent profile $q \in \Pi_{-i}$, let $B_i(q) \subset \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$ denote the set of stationary distributions supported entirely on Bellman-maximizing actions at every state:

$$B_i(q) := \left\{ \pi_i \in \Pi_i \;\middle|\; \forall s \in \mathcal{S},\ \mathrm{supp}(\pi_i(\cdot \mid s)) \subseteq \arg\max_{a_i \in \mathcal{A}_i} (\mathcal{T}_i^{a_i,q} W_i^q)(s) \right\}.$$

By Lemma C.1, $B_i(q)$ is non-empty, convex, and compact. Upper hemicontinuity of the argmax correspondence ensures that the product correspondence $B(\pi) := \prod_{i \in \mathcal{N}} B_i(\pi_{-i})$ maps the compact convex set Π into non-empty compact convex subsets of itself with a closed graph. By Kakutani's fixed-point theorem, there exists $\pi^* \in B(\pi^*)$, which constitutes a stationary Nash equilibrium. When $K \geq |\mathcal{S}|$, the identity mapping $\phi(s) = s$ induces $\Pi_\phi = \Pi$, whence $J_{|\mathcal{S}|}(\mathcal{G}, \mathcal{Q}) = 0$.

C.3 Best-Response Loss as an Upper Bound on Restricted Unilateral Gain

For an initial state distribution $\mu \in \Delta(\mathcal{S})$, a baseline abstract policy $p \in \Pi_{\phi,i}$, and an opponent profile $q \in \Pi_{-i}$, define the unrestricted unilateral deviation gain and the restricted deviation gain within Π_ϕ as:

$$D_q(p) := \mu^\top W_i^q - \mu^\top V_i^{p,q}, \qquad D_{\phi,q}(p) := \sup_{\sigma \in \Pi_{\phi,i}} \mu^\top V_i^{\sigma,q} - \mu^\top V_i^{p,q}.$$

Subtracting the restricted gain from the unrestricted gain cancels the baseline value $\mu^\top V_i^{p,q}$:

$$D_q(p) - D_{\phi,q}(p) = \mu^\top W_i^q - \sup_{\sigma \in \Pi_{\phi,i}} \mu^\top V_i^{\sigma,q} = \inf_{\sigma \in \Pi_{\phi,i}} \mu^\top (W_i^q - V_i^{\sigma,q}).$$

By Lemma C.1, $W_i^q(s) - V_i^{\sigma,q}(s) \geq 0$ for all $s \in \mathcal{S}$. Consequently,

$$0 \leq \mu^\top (W_i^q - V_i^{\sigma,q}) \leq \|W_i^q - V_i^{\sigma,q}\|_\infty.$$

Taking the infimum over $\sigma \in \Pi_{\phi,i}$ yields

$$0 \leq D_q(p) - D_{\phi,q}(p) \leq S_q(\phi). \tag{15}$$

Equation (15) demonstrates that the best-response loss $S_q(\phi)$ provides an explicit upper bound on the strategic suboptimality introduced by restricting counter-strategies to the abstract class Π_ϕ.

C.4 Minimax Decomposition of the Joint Objective

For a fixed partition mapping $\phi : \mathcal{S} \to [K]$, the equilibrium transfer error and the opponent-conditioned best-response loss are defined respectively as

$$E_\phi(\mathcal{G}) = \inf_{\pi \in \Pi_\phi} \mathrm{Gap}_\mathcal{G}(\pi), \qquad S_\mathcal{Q}(\phi) = \max_{q \in \mathcal{Q}} \inf_{\sigma_q \in \Pi_{\phi,i(q)}} \|W_{i(q)}^q - V_{i(q)}^{\sigma_q,q}\|_\infty.$$

Because the deployed profile π and the collection of response heads $\{\sigma_q\}_{q \in \mathcal{Q}}$ do not share decision variables beyond the common state partition ϕ, the joint optimization decouples:

$$\begin{aligned} &\min_{\substack{\pi \in \Pi_\phi \\ \sigma_q \in \Pi_{\phi,i(q)}\ (q \in \mathcal{Q})}} \max\left\{ \mathrm{Gap}_\mathcal{G}(\pi), \max_{q \in \mathcal{Q}} \|W_{i(q)}^q - V_{i(q)}^{\sigma_q,q}\|_\infty \right\} \\ &= \max\left\{ \min_{\pi \in \Pi_\phi} \mathrm{Gap}_\mathcal{G}(\pi), \max_{q \in \mathcal{Q}} \min_{\sigma_q \in \Pi_{\phi,i(q)}} \|W_{i(q)}^q - V_{i(q)}^{\sigma_q,q}\|_\infty \right\} \\ &= \max\{E_\phi(\mathcal{G}), S_\mathcal{Q}(\phi)\}. \end{aligned} \tag{16}$$

Minimizing Equation (16) over all admissible state partitions with $|\phi(\mathcal{S})| \leq K$ yields the joint abstraction objective $J_K(\mathcal{G}, \mathcal{Q})$ in Equation (4).

C.5 Exactness of the Joint Bellman Program

We now prove that the joint Bellman program defined in Equation (5) is exact.

Lemma C.2 (Exactness of the Joint Bellman Program)**.** *The global optimum t^* of the mathematical program* (5)*, optimized over partition mappings $\phi : \mathcal{S} \to [K]$, ϕ-abstract deployment policies $\pi \in \Pi_\phi$, and ϕ-abstract response heads $\sigma_q \in \Pi_{\phi,i(q)}$ $(q \in \mathcal{Q})$, equals $J_K(\mathcal{G}, \mathcal{Q})$.*

Proof. We establish the equality by proving both directional inequalities.

Part 1: Optimum $\geq J_K(\mathcal{G}, \mathcal{Q})$. Let $(\phi, \pi, \{\sigma_q\}, \{v_i, u_i\}_{i\in\mathcal{N}}, \{v_q, u_q\}_{q\in\mathcal{Q}}, t)$ be an arbitrary feasible assignment for program (5). For each player $i \in \mathcal{N}$, the evaluation constraint $v_i = \mathcal{T}_i^{\pi} v_i$ together with fixed-point uniqueness implies $v_i = V_i^{\pi}$. The deviation inequality $u_i \geq \mathcal{T}_i^{a_i,\pi_{-i}} u_i$ holding for all $a_i \in \mathcal{A}_i$ implies

$$u_i \geq \max_{a_i\in\mathcal{A}_i} \mathcal{T}_i^{a_i,\pi_{-i}} u_i = \mathcal{T}_i^{*,\pi_{-i}} u_i.$$

By Lemma C.1, $u_i \geq W_i^{\pi_{-i}}$ coordinatewise. The slack constraint $u_i - v_i \leq t\mathbf{1}$ then ensures

$$W_i^{\pi_{-i}} - V_i^{\pi} \leq u_i - v_i \leq t\mathbf{1},$$

whence $\mathrm{Gap}_{\mathcal{G}}(\pi) \leq t$.

Similarly, for each benchmark opponent $q \in \mathcal{Q}$ facing player i, the constraints $v_q = \mathcal{T}_i^{\sigma_q,q} v_q$ and $u_q \geq \mathcal{T}_i^{a_i,q} u_q$ for all $a_i \in \mathcal{A}_i$ imply $v_q = V_i^{\sigma_q,q}$ and $u_q \geq W_i^q$. Consequently,

$$0 \leq W_i^q - V_i^{\sigma_q,q} \leq u_q - v_q \leq t\mathbf{1},$$

which implies $S_q(\phi) \leq \|W_i^q - V_i^{\sigma_q,q}\|_\infty \leq t$ for all $q \in \mathcal{Q}$. Because π and all σ_q are ϕ-abstract, this feasible solution implies $J_K(\mathcal{G}, \mathcal{Q}) \leq \max\{E_\phi(\mathcal{G}), S_{\mathcal{Q}}(\phi)\} \leq t$. Since this bound holds for any feasible candidate, $t^* \geq J_K(\mathcal{G}, \mathcal{Q})$.

Part 2: Optimum $\leq J_K(\mathcal{G}, \mathcal{Q})$. By the attainment established in Lemma C.1, let $(\phi^*, \pi^*, \{\sigma_q^*\})$ attain the joint optimum $J_K(\mathcal{G}, \mathcal{Q})$. For each player $i \in \mathcal{N}$ and task $q \in \mathcal{Q}$, define the witness vectors:

$$v_i = V_i^{\pi^*}, \qquad u_i = W_i^{\pi^*_{-i}}, \qquad v_q = V_i^{\sigma_q^*,q}, \qquad u_q = W_i^q.$$

By Bellman fixed-point definitions, $v_i = \mathcal{T}_i^{\pi^*} v_i$ and $v_q = \mathcal{T}_i^{\sigma_q^*,q} v_q$. Furthermore, because $u_i = W_i^{\pi^*_{-i}} = \mathcal{T}_i^{*,\pi^*_{-i}} u_i$, it satisfies $u_i \geq \mathcal{T}_i^{\pi^*_{-i},a_i} u_i$ for all $a_i \in \mathcal{A}_i$. Analogously, $u_q \geq \mathcal{T}_i^{a_i,q} u_q$ for all $a_i \in \mathcal{A}_i$.

By definition of the joint objective value $J_K(\mathcal{G}, \mathcal{Q})$, we have:

$$u_i - v_i = W_i^{\pi^*_{-i}} - V_i^{\pi^*} \leq \mathrm{Gap}_{\mathcal{G}}(\pi^*)\mathbf{1} \leq J_K(\mathcal{G}, \mathcal{Q})\mathbf{1},$$

$$u_q - v_q = W_i^q - V_i^{\sigma_q^*,q} \leq S_q(\phi^*)\mathbf{1} \leq J_K(\mathcal{G}, \mathcal{Q})\mathbf{1}.$$

Setting $t = J_K(\mathcal{G}, \mathcal{Q})$ yields a fully feasible assignment for program (5), proving that $t^* \leq J_K(\mathcal{G}, \mathcal{Q})$. Combining both inequalities establishes $t^* = J_K(\mathcal{G}, \mathcal{Q})$. □

C.6 Treatment of State-Dependent Legal Actions

In games with state-dependent legal action constraints $\mathcal{A}_i(s) \subseteq \mathcal{A}_i$, the Bellman optimality operators are restricted to legal actions:

$$(\mathcal{T}_i^{*,q} v)(s) := \max_{a_i\in\mathcal{A}_i(s)} (\mathcal{T}_i^{a_i,q} v)(s).$$

Any admissible state partition ϕ must be compatible with the action masks, satisfying $\mathcal{A}_i(s) = \mathcal{A}_i(s')$ whenever $\phi(s) = \phi(s')$. For all illegal actions $a_i \notin \mathcal{A}_i(s)$, policy probabilities are constrained to zero: $\pi_i(a_i \mid s) = 0$. This preserves the validity of Lemma C.2 and ensures that decoded abstract policies remain strictly legal in the underlying game.

D EXPONENTIAL SEPARATION BETWEEN DECISION AND VALUE ABSTRACTION

This appendix provides the complete proof of Theorem 4.1, establishing an exponential separation between the state partition cardinality required for strategy completeness versus value reconstruction under identical benchmark tasks. In addition, we present the formal analysis of response conflicts, graph-theoretic hardness, and the motivating counterexamples.

D.1 PROOF OF THEOREM 4.1 (EXPONENTIAL SEPARATION)

Game Construction. Let $m \geq 1$ and $n_0 \geq 0$ be integers. Set $D := m + n_0$, and define the finite state space as the binary hypercube:

$$\mathcal{S} = \{0,1\}^m \times \{0,1\}^{n_0} = \{0,1\}^D.$$

Each state is represented by coordinates $s = z = (x, y)$, where $x = (x_1, \ldots, x_m) \in \{0,1\}^m$ denotes the policy-relevant coordinates and $y = (y_1, \ldots, y_{n_0}) \in \{0,1\}^{n_0}$ denotes the payoff-shifting coordinates.

Let P be an arbitrary joint-action transition kernel on $\mathcal{S}$, and let $\gamma \in (0,1)$ be the discount factor with effective horizon $H = (1-\gamma)^{-1}$. The game is a two-player constant-sum game. Player 1 (the row player) has action space $\mathcal{A}_1 = \{0,1\}$. Player 2 (the column player) has a neutral action $\perp$ and task actions indexed by triples:

$$q = (j, k, c) \in [m] \times [D] \times \{0,1\}.$$

For a fixed non-empty coordinate subset $T \subseteq [m]$, the benchmark opponent family is defined as:

$$\mathcal{Q}_T := \{q = (j,k,c) \mid j \in T,\ k \in [D],\ c \in \{0,1\}\}.$$

We define the scaling constants:

$$\nu = \frac{1}{16}, \qquad w_j \in \left[\frac{1}{8}, \frac{3}{16}\right] \quad (\forall j \in [m]).$$

For each task $q = (j,k,c)$, we specify the target optimal action $a_q^*(s)$ and target value function $U_q(s)$ as:

$$a_q^*(s) := x_j \oplus c, \qquad U_q(s) := H\left(\frac{1}{2} + w_j\right) + \nu z_k. \tag{17}$$

For a given state $s \in \mathcal{S}$ and column action q, let $P_m(\cdot \mid s) := P(\cdot \mid s, a_q^*(s), q)$ denote the transition probability vector under the matching action $a_q^*(s)$, and let $P_w(\cdot \mid s) := P(\cdot \mid s, 1 - a_q^*(s), q)$ denote the transition vector under the non-matching (wrong) action.

We define the stage rewards for Player 1 under column action q by:

$$r_m(s) := U_q(s) - \gamma(P_m U_q)(s), \tag{18}$$

$$r_w(s) := 1 - r_m(s). \tag{19}$$

Under the neutral column action $\perp$, both actions $a_1 \in \{0,1\}$ yield stage reward $r_1(s, a_1, \perp) = 1/2$. Player 2 receives complementary reward $r_2 = 1 - r_1$, ensuring a constant-sum game with stage-sum 1.

Reward Boundedness. Using $P_m \mathbf{1} = \mathbf{1}$ and $H(1-\gamma) = 1$, expanding Equation (18) gives:

$$\begin{aligned} r_m(s) &= H\left(\frac{1}{2} + w_j\right) + \nu z_k - \gamma H\left(\frac{1}{2} + w_j\right) - \gamma\nu(P_m z_k)(s) \\ &= \frac{1}{2} + w_j + \nu\left(z_k - \gamma(P_m z_k)(s)\right). \end{aligned} \tag{20}$$

Because $z_k \in \{0,1\}$ and $(P_m z_k)(s) \in [0,1]$, the term satisfies $-\gamma \leq z_k - \gamma(P_m z_k)(s) \leq 1$. Given $\nu = 1/16$ and $w_j \in [1/8, 3/16]$, we have:

$$\frac{1}{2} + \frac{1}{8} - \frac{1}{16} \leq r_m(s) \leq \frac{1}{2} + \frac{3}{16} + \frac{1}{16} \implies r_m(s) \in \left[\frac{9}{16}, \frac{3}{4}\right].$$

Consequently, $r_w(s) = 1 - r_m(s) \in [1/4, 7/16]$. Thus, all stage rewards strictly reside in $[0,1]$.

Optimality of the Matching Action and Bellman Deficit. We now verify that U_q is the unique unrestricted optimal value vector W_1^q for task q. For the matching action $a_q^*(s)$, Equation (18) yields:

$$r_m(s) + \gamma(P_m U_q)(s) = U_q(s).$$

For the non-matching action $1 - a_q^*(s)$, using $r_w = 1 - r_m$:

$$\begin{aligned} U_q(s) - [r_w(s) + \gamma(P_w U_q)(s)] &= U_q(s) - 1 + r_m(s) - \gamma(P_w U_q)(s) \\ &= 2U_q(s) - 1 - \gamma(P_m U_q)(s) - \gamma(P_w U_q)(s). \end{aligned} \tag{21}$$

Substituting $U_q(s) = H(1/2 + w_j) + \nu z_k$ into (21), the constant component simplifies to:

$$2H\left(\frac{1}{2} + w_j\right) - 1 - 2\gamma H\left(\frac{1}{2} + w_j\right) = 2(1-\gamma)H\left(\frac{1}{2} + w_j\right) - 1 = 2w_j.$$

The state-dependent component is given by $\nu[2z_k - \gamma(P_m + P_w)z_k]$. Since expectations lie in $[0, 1]$,

$$2z_k - \gamma(P_m + P_w)z_k \geq -2\gamma.$$

Therefore, the Bellman deficit $g_q(s)$ of the sub-optimal action satisfies:

$$g_q(s) := U_q(s) - [r_w(s) + \gamma(P_w U_q)(s)] \geq 2w_j - 2\gamma\nu = 2(w_j - \gamma\nu) > 0. \tag{22}$$

Because $g_q(s) > 0$ uniformly, $a_q^*(s)$ is the unique Bellman-maximizing action at every state s. Thus, $\mathcal{T}_q^* U_q = U_q$, which by Lemma C.1 implies:

$$W_1^q = U_q. \tag{23}$$

For any stationary response policy $\sigma \in \Pi_1$, define the one-step Bellman deficit vector $d_q^\sigma := W_1^q - \mathcal{T}_1^{\sigma,q} W_1^q \geq 0$. Since $V_1^{\sigma,q} = \mathcal{T}_1^{\sigma,q} V_1^{\sigma,q}$, subtracting this identity gives:

$$W_1^q - V_1^{\sigma,q} = d_q^\sigma + \gamma P_1^{\sigma,q}(W_1^q - V_1^{\sigma,q}) = (I - \gamma P_1^{\sigma,q})^{-1} d_q^\sigma = \sum_{t=0}^{\infty} \gamma^t (P_1^{\sigma,q})^t d_q^\sigma.$$

Because all entries of $(P_1^{\sigma,q})^t$ and d_q^σ are nonnegative, the $t = 0$ identity term yields:

$$W_1^q - V_1^{\sigma,q} \geq d_q^\sigma \quad \text{coordinatewise.} \tag{24}$$

Furthermore, since $\|(I - \gamma P_1^{\sigma,q})^{-1}\|_\infty = H$, we obtain the upper bound:

$$\|W_1^q - V_1^{\sigma,q}\|_\infty \leq H \|d_q^\sigma\|_\infty. \tag{25}$$

Part 1: Best-Response Abstraction Bound ($K_{\mathrm{BR}}(\mathcal{Q}_T) = 2^{|T|}$). We establish that achieving best-response error $S_{\mathcal{Q}_T}(\phi) \leq \varepsilon_0 = 1/64$ requires precisely $2^{|T|}$ abstract states.

Upper bound ($K_{\mathrm{BR}}(\mathcal{Q}_T) \leq 2^{|T|}$): Define partition ϕ such that $\phi(s) = \phi(s')$ if and only if $x_T = x'_T$, where $x_T := (x_j)_{j \in T}$. This partition uses exactly $2^{|T|}$ abstract states. For any task $q = (j, k, c) \in \mathcal{Q}_T$, all states in a partition cell share the identical coordinate x_j, and thus require the identical matching action $a_q^*(s) = x_j \oplus c$. Assigning this pure action to the abstract response policy on each cell achieves $d_q^\sigma(s) = 0$ for all $s \in \mathcal{S}$, which by Equation (25) implies:

$$S_{\mathcal{Q}_T}(\phi) = 0 < \varepsilon_0.$$

Lower bound ($K_{\mathrm{BR}}(\mathcal{Q}_T) \geq 2^{|T|}$): Suppose an encoder ϕ uses fewer than $2^{|T|}$ abstract states. Since there are $2^{|T|}$ distinct binary patterns on x_T, by the pigeonhole principle there exists at least one partition cell containing two states s, s' with $x_T \neq x'_T$. Choose an index $j \in T$ such that $x_j \neq x'_j$. For any fixed $k \in [D]$ and $c \in \{0, 1\}$, consider task $q = (j, k, c) \in \mathcal{Q}_T$. The required optimal actions satisfy $a_q^*(s) \neq a_q^*(s')$.

Any ϕ-abstract policy $\sigma \in \Pi_{\phi,1}$ must assign a state-independent action distribution on this cell. Let $p \in [0, 1]$ denote the probability assigned to action 0. At one of the two states s, s', the probability of selecting the sub-optimal action is at least $\max\{p, 1 - p\} \geq 1/2$. At that state, the one-step Bellman deficit satisfies:

$$d_q^\sigma \geq \frac{1}{2} g_q \geq w_j - \gamma\nu.$$

Using $w_j \geq 1/8$, $\nu = 1/16$, and $\gamma < 1$:

$$w_j - \gamma\nu > \frac{1}{8} - \frac{1}{16} = \frac{1}{16}.$$

Applying Equation (24) yields:

$$\|W_1^q - V_1^{\sigma,q}\|_\infty \geq \frac{1}{16} > \frac{1}{64} = \varepsilon_0.$$

Since this holds for every abstract policy $\sigma \in \Pi_{\phi,1}$, any partition achieving $S_{\mathcal{Q}_T}(\phi) \leq \varepsilon_0$ must separate all $2^{|T|}$ configurations on x_T. Thus, $K_{\mathrm{BR}}(\mathcal{Q}_T) = 2^{|T|}$.

Part 2: Value Reconstruction Bound ($K_{\mathrm{value}}(\mathcal{Q}_T) = 2^{m+n_0}$). Now consider the minimum partition cardinality K_{value} required by any encoder $\phi : \mathcal{S} \to [K]$ paired with a task-specific scalar decoder $v_q : [K] \to \mathbb{R}$ to uniformly reconstruct optimal values: $\max_{q\in\mathcal{Q}_T} \|W_1^q - v_q \circ \phi\|_\infty \leq \varepsilon_0$.

Fix any distinct pair of states $s \neq s' \in \mathcal{S}$. Then their full coordinates differ: $z \neq z'$. Thus, there exists a coordinate index $k \in [D]$ such that $z_k \neq z'_k$. Since T is non-empty, fix $j \in T$ and $c \in \{0, 1\}$, and select task $q = (j, k, c) \in \mathcal{Q}_T$. By Equation (23):

$$|W_1^q(s) - W_1^q(s')| = \nu|z_k - z'_k| = \nu = \frac{1}{16}. \tag{26}$$

If s and s' share an abstract state $\phi(s) = \phi(s')$, any scalar decoder must output an identical value $v_q(\phi(s)) = v_q(\phi(s'))$. By the triangle inequality:

$$|W_1^q(s) - W_1^q(s')| \leq |W_1^q(s) - v_q(\phi(s))| + |v_q(\phi(s')) - W_1^q(s')|.$$

Consequently, at least one of the states incurs an error lower-bounded by:

$$\max\left\{|W_1^q(s) - v_q(\phi(s))|,\ |W_1^q(s') - v_q(\phi(s'))|\right\} \geq \frac{\nu}{2} = \frac{1}{32} > \varepsilon_0.$$

Therefore, no two distinct states in $\mathcal{S}$ can be merged. Reconstructing the optimal values of $\mathcal{Q}_T$ within error ε_0 requires:

$$K_{\mathrm{value}}(\mathcal{Q}_T) \geq |\mathcal{S}| = 2^{m+n_0}.$$

Assigning a unique abstract state to every original state achieves zero error, proving $K_{\mathrm{value}}(\mathcal{Q}_T) = 2^{m+n_0}$.

Part 3: Single-State Equilibrium Profile ($K_{\mathrm{equilibrium}} = 1$). Consider the candidate strategy profile where Player 1 plays the uniform mixed policy $\pi_1(0 \mid s) = \pi_1(1 \mid s) = 1/2$ at all states, and Player 2 deterministically plays the neutral action $\pi_2(\perp \mid s) = 1$ at all states.

Against $\perp$, both of Player 1's actions yield immediate reward $1/2$. Setting constant continuation value $c := H/2 = \frac{1}{2(1-\gamma)}$, for any action $a_1 \in \{0, 1\}$:

$$\frac{1}{2} + \gamma \sum_{s'\in\mathcal{S}} P(s' \mid s, a_1, \perp)c = \frac{1}{2} + \gamma c = \frac{1}{2} + \frac{\gamma}{2(1-\gamma)} = \frac{1}{2(1-\gamma)} = c.$$

Thus, the constant vector $c\mathbf{1}$ satisfies the Bellman optimality equation against $\perp$, and Player 1 has zero incentive to deviate.

Conversely, fix any column action $q = (j, k, c') \in \mathcal{Q}_T$. Under Player 1's uniform strategy, the expected immediate reward for Player 1 is $\frac{1}{2}r_m(s) + \frac{1}{2}r_w(s) = 1/2$ by Equation (19). Player 2's complementary immediate reward is therefore identically $1/2$ across all states. The continuation value of $c\mathbf{1}$ evaluates to γc regardless of the transition row $P(\cdot \mid s, a_1, q)$. Hence, every action available to Player 2 yields return c, leaving zero profitable deviation from $\perp$. The profile is a stationary Nash equilibrium supported on a single abstract state:

$$K_{\mathrm{equilibrium}} = 1.$$

Part 4: Robustness to Perturbations. Let $\widetilde{\mathcal{G}}$ be a perturbed game satisfying $|\widetilde{r}_i(s,\mathbf{a}) - r_i(s,\mathbf{a})| \le \zeta$ and $\mathrm{TV}(\widetilde{P}(\cdot \mid s,\mathbf{a}), P(\cdot \mid s,\mathbf{a})) \le \eta$ uniformly. For any bounded value vector $v \in [0,H]^{|\mathcal{S}|}$,

$$\left| \sum_{s'} (P(s' \mid s,\mathbf{a}) - \widetilde{P}(s' \mid s,\mathbf{a}))v(s') \right| \le \mathrm{TV}(P,\widetilde{P})\,\mathrm{span}(v) \le \eta H.$$

For any stationary policy τ, the value difference satisfies:

$$\|V_i^\tau - \widetilde{V}_i^\tau\|_\infty \le \zeta + \gamma\|P^\tau(V_i^\tau - \widetilde{V}_i^\tau)\|_\infty + \gamma\|(P^\tau - \widetilde{P}^\tau)\widetilde{V}_i^\tau\|_\infty \le \zeta + \gamma\|V_i^\tau - \widetilde{V}_i^\tau\|_\infty + \gamma\eta H.$$

Rearranging yields the uniform value perturbation bound:

$$\|V_i^\tau - \widetilde{V}_i^\tau\|_\infty \le H\zeta + \gamma H^2\eta =: e. \tag{27}$$

By contraction arguments, the optimal response values satisfy $\|W_i^q - \widetilde{W}_i^q\|_\infty \le e$. By the triangle inequality, for any partition ϕ and response σ:

$$\left| \|W_i^q - V_i^{\sigma,q}\|_\infty - \|\widetilde{W}_i^q - \widetilde{V}_i^{\sigma,q}\|_\infty \right| \le 2e.$$

Choosing perturbation parameters such that $e \le 1/256$, the loss of the $2^{|T|}$-state partition in $\widetilde{\mathcal{G}}$ is at most $2e \le 1/128 < \varepsilon_0$. Conversely, any partition with fewer than $2^{|T|}$ states incurs loss in $\widetilde{\mathcal{G}}$ of at least $1/16 - 2e \ge 7/128 > \varepsilon_0$. Similarly, the value separation between distinct states in $\widetilde{\mathcal{G}}$ satisfies $|W_1^q(s) - W_1^q(s')| \ge 1/16 - 2e \ge 7/128 > 2\varepsilon_0 = 1/32$, preserving $K_{\mathrm{value}}(\mathcal{Q}_T) = 2^{m+n_0}$. This completes the proof of Theorem 4.1. □

D.2 Response Conflicts and Graph-Theoretic Characterization

Lemma D.1 (Common Optimal Actions across Cells)**.** *For a finite task family $\mathcal{Q}$, $S_{\mathcal{Q}}(\phi) = 0$ if and only if, for every task $q \in \mathcal{Q}$ and every non-empty partition cell $B = \phi^{-1}(k)$, the state-level optimal action sets share a non-empty intersection:*

$$\bigcap_{s\in B} A_q^*(s) \neq \emptyset, \qquad \textit{where } A_q^*(s) := \arg\max_{a_i \in \mathcal{A}_i} (\mathcal{T}_i^{a_i,q} W_i^q)(s).$$

Proof. If the intersection is non-empty, select an action $a_{B,q} \in \bigcap_{s\in B} A_q^*(s)$ for each cell B, and define the abstract response policy as $\sigma_q(a_{B,q} \mid s) = 1$ for all $s \in B$. Then for all $s \in \mathcal{S}$, $(\mathcal{T}_i^{\sigma_q,q} W_i^q)(s) = (\mathcal{T}_i^{*,q} W_i^q)(s) = W_i^q(s)$. By fixed-point uniqueness, $V_i^{\sigma_q,q} = W_i^q$, whence $S_{\mathcal{Q}}(\phi) = 0$.

Conversely, assume $S_{\mathcal{Q}}(\phi) = 0$. There exists an abstract policy $\sigma \in \Pi_{\phi,i}$ such that $V_i^{\sigma,q} = W_i^q$. By Bellman optimality,

$$0 = W_i^q(s) - (\mathcal{T}_i^{\sigma,q} W_i^q)(s) = \sum_{a_i \in \mathcal{A}_i} \sigma(a_i \mid \phi(s)) \left[W_i^q(s) - (\mathcal{T}_i^{a_i,q} W_i^q)(s) \right].$$

Because each term $W_i^q(s) - (\mathcal{T}_i^{a_i,q} W_i^q)(s) \ge 0$, a convex combination can be zero only if every action in the support of $\sigma(\cdot \mid \phi(s))$ achieves zero deficit at every state $s \in B$. Since the support of a probability distribution is non-empty, the intersection $\bigcap_{s\in B} A_q^*(s)$ is non-empty. □

Proposition D.2 (Conflict Graph Proper Coloring and One-Step Deficit Bounds)**.** *Assume binary action spaces $|\mathcal{A}_i| = 2$. Define the conflict graph $G_{\mathrm{C}} = (\mathcal{S}, \mathcal{E}_{\mathrm{C}})$ by placing an undirected edge between s and t whenever there exists a task $q \in \mathcal{Q}$ such that $A_q^*(s) \cap A_q^*(t) = \emptyset$. Then $S_{\mathcal{Q}}(\phi) = 0$ if and only if ϕ is a proper vertex coloring of G_{C}.*

Furthermore, let $g_q(s)$ denote the Bellman deficit of the unique sub-optimal action at state s for task q. For opposite unique preferences between states s and t, define the harmonic conflict weight:

$$w_q(s,t) := \frac{g_q(s) g_q(t)}{g_q(s) + g_q(t)}, \qquad w(s,t) := \max_{q\in\mathcal{Q}} w_q(s,t).$$

Set $w_q(s,t) = 0$ when the pair has no opposite unique preferences for task q. Then the maximum monochromatic edge weight $b_{\mathcal{Q}}(\phi) := \max_{s,t:\phi(s)=\phi(t)} w(s,t)$ satisfies:

$$b_{\mathcal{Q}}(\phi) \le S_{\mathcal{Q}}(\phi) \le H b_{\mathcal{Q}}(\phi). \tag{28}$$

Proof. For binary actions $\{0, 1\}$, two non-empty sets have an empty intersection if and only if one is $\{0\}$ and the other is $\{1\}$. Lemma D.1 directly implies that zero loss corresponds to proper graph coloring.

For the quantitative bound, consider task q on a partition cell B containing states s and t with conflicting preferences ($A_q^*(s) = \{0\}$ and $A_q^*(t) = \{1\}$). Let $p \in [0, 1]$ be the cell's probability of choosing action 0. The one-step deficits at s and t are $g_q(s)(1 - p)$ and $g_q(t)p$, respectively. Minimizing the maximum one-step deficit over $p \in [0, 1]$ gives:

$$\min_{p\in[0,1]} \max\{g_q(s)(1-p),\ g_q(t)p\} = \frac{g_q(s)g_q(t)}{g_q(s) + g_q(t)} = w_q(s, t).$$

Since the resolvent satisfies $\|(I - \gamma P_i^{\sigma,q})^{-1} d_q^\sigma\|_\infty \geq \|d_q^\sigma\|_\infty$, minimizing over abstract policies yields $S_{\mathcal{Q}}(\phi) \geq b_{\mathcal{Q}}(\phi)$. Conversely, selecting the policy that minimizes the one-step deficit on each cell guarantees $\|d_q^\sigma\|_\infty \leq b_{\mathcal{Q}}(\phi)$, and applying the resolvent bound $\|(I - \gamma P_i^{\sigma,q})^{-1}\|_\infty = H$ yields $S_{\mathcal{Q}}(\phi) \leq H b_{\mathcal{Q}}(\phi)$. □

Proposition D.3 (NP-Completeness of Zero Best-Response Loss)**.** *Deciding whether $D_K(\mathcal{G}, \mathcal{Q}) = 0$ is NP-complete for explicitly represented finite task games. Hardness holds for $K = 3$, binary focal actions, self-loop transitions, and games possessing a single-state equilibrium.*

Proof. Membership in NP. Given a candidate partition $\phi : \mathcal{S} \to [K]$, verifying $S_{\mathcal{Q}}(\phi) = 0$ requires evaluating Bellman optimality sets $A_q^*(s)$ via linear programming for each task $q \in \mathcal{Q}$ and checking whether $\bigcap_{s\in B} A_q^*(s) \neq \emptyset$ for each cell B, which executes in polynomial time.

NP-hardness. We reduce from graph 3-colorability (Karp, 1972). Given an undirected graph $G_0 = (V, E)$, construct a state space with one state per vertex $s_u \in \mathcal{S}$ for each $u \in V$. All states self-loop under transition kernel $P(s_u \mid s_u, \mathbf{a}) = 1$ with discount $\gamma = 1/2$. For each edge $e = (u, v) \in E$, introduce an opponent task q_e such that action 0 is uniquely optimal at s_u with deficit $1/4$, action 1 is uniquely optimal at s_v with deficit $1/4$, and both actions are tied at all other states. By Proposition D.2, the conflict graph G_{C} is isomorphic to G_0. Concretely, give the preferred action reward $5/8$ and the other action $3/8$ at each edge endpoint, and give both actions $1/2$ elsewhere. Add a neutral opponent action paying $1/2$ for either focal action, and complementary rewards for the opponent. Uniform focal play and the neutral opponent action form a single-state equilibrium. A 3-partition achieves $S_{\mathcal{Q}}(\phi) = 0$ if and only if G_0 admits a proper 3-coloring. □

D.3 Exact Trade-off for the Self-Loop Tactical Family

Consider the self-loop game family where each state $s = (x, y) \in \{0, 1\}^m \times \{0, 1\}^{n_0}$ has stage reward

$$r((x, y), a, (j, c)) = \frac{1}{2} + g(y) + w_j(2\mathbf{1}\{a = x_j \oplus c\} - 1),$$

with ordered weights $w_1 > w_2 > \cdots > w_m > 0$. Merging states differing in coordinate x_j incurs a one-step stage loss of w_j. Under self-loop transitions, stage losses accumulate by the factor H. Therefore, at target tolerance ε, every coordinate satisfying $Hw_j > \varepsilon$ must be distinguished. Consequently, the minimum partition cardinality and optimal loss satisfy:

$$K_{\mathcal{Q}_T}(\varepsilon) = 2^{|\{j\in T: Hw_j > \varepsilon\}|}, \qquad D_K(\mathcal{G}, \mathcal{Q}_{[m]}) = Hw_{1+\lfloor \log_2 K \rfloor} \quad (\forall K < 2^m). \tag{29}$$

D.4 Analytical Construction of Motivating Counterexamples

Counterexample 1: Zero Prediction Error with $\Omega(H)$ Best-Response Regret. Let $\mathcal{S} = \{s_0, s_1\}$ be sampled independently and uniformly at each step. Player 1 has actions $\{\mathrm{safe}, \mathrm{left}, \mathrm{right}\}$ with rewards:

$$r(s_0, \mathrm{safe}) = 1/2, \quad r(s_0, \mathrm{left}) = 1, \quad r(s_0, \mathrm{right}) = 0,$$
$$r(s_1, \mathrm{safe}) = 1/2, \quad r(s_1, \mathrm{left}) = 0, \quad r(s_1, \mathrm{right}) = 1.$$

Under the behavioral policy choosing safe, the reward is identically $1/2$ and next-state transitions are uniformly distributed on $\{s_0, s_1\}$ regardless of action. Transition and reward prediction models trained on-policy incur strictly zero error. Moreover, any single-state abstract policy parameterized

by action probabilities (p_s, p_l, p_r) has expected stage reward $\frac{1}{2}(p_s/2 + p_l) + \frac{1}{2}(p_s/2 + p_r) = 1/2$, resulting in discounted return $H/2$ from the uniform initial distribution. The safe policy has value $H/2$ at both states, and no abstract policy has a larger minimum state value. However, an unrestricted optimal response selects left at s_0 and right at s_1, achieving constant reward 1 and discounted value H. The abstract best-response loss is therefore $H - H/2 = H/2 = \Omega(H)$.

Counterexample 2: Exact Single-State Equilibrium with $\Omega(H)$ Best-Response Regret. Consider a two-state matrix game with states $\{s_0, s_1\}$ chosen uniformly at each step, with payoff matrices:

$$R_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad R_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and complementary column player payoffs. Under uniform play by both players, the expected row payoff is $1/2$ across all actions and states, yielding constant Bellman value $H/2$ and constituting an exact single-state abstract equilibrium. Against a fixed opponent choosing column action 0, the unique best response requires playing row 0 at s_0 and row 1 at s_1. Every single-state policy has uniform-initial-state return $H/2$, so its worst-state response loss is at least $H/2$. The uniform action mixture attains value $H/2$ at both states, giving $S_q(\phi) = H/2$.

E Bounded-treewidth joint optimization

This appendix proves Theorem 5.1 for the joint objective in Equation 4. The foundational Bellman facts and the proof that Equation 5 has optimum J_K are supplied in Appendix C. Those facts use finite discounted games and do not impose a transition factorization.

E.1 Variables and the canonical finite problem

Write $N = |\mathcal{S}|$ and take $1 \le K \le N$ without loss of generality. For each deployment player and each declared response task, create a head index g with action alphabet of size A_g. A deployment player has one head in its own mixed product profile; a task head faces fixed opponents. Introduce:

1. a state label $z_s \in [K]$ for every s;
2. code probabilities $x_{k,g,a}$ and local probabilities $p_{s,g,a}$, for $a = 1, \dots, A_g - 1$;
3. normalized evaluation and optimal-response witnesses $v_g(s), u_g(s) \in [0, 1]$.

The last probability is $1 - \sum_{a<A_g} p_{s,g,a}$ or its code counterpart. Require nonnegativity of all listed coordinates and that every such partial sum is at most one. For every s, k, g, a, the selector constraint is

$$z_s \neq k \ \vee \ p_{s,g,a} = x_{k,g,a}. \tag{30}$$

Exactly one z_s is selected. Consequently all local probabilities on the same code equal the same code probabilities, including the last action. Empty cells are permitted. Conversely every code policy admits this representation by copying its probabilities to the appropriate local variables. Selectors introduce no extra state information at deployment.

For a vector $y \in [0, 1]^N$, define the normalized one-step operator

$$(\mathcal{B}_i^{\mathbf{a}} y)(s) = (1 - \gamma) r_i(s, \mathbf{a}) + \gamma \sum_{s'} P(s' \mid s, \mathbf{a}) y(s').$$

Averaging over the relevant stationary product actors defines $\mathcal{B}_i^{\pi}$ and $\mathcal{B}_i^{a_i, \pi_{-i}}$. Averaging the declared opponents defines $\mathcal{B}_h^{a}$ and $\mathcal{B}_h^{\sigma_h}$. Each is an expectation of a variable in $[0, 1]$. The code constraints and these operators are precisely the normalized counterpart of Equation 5.

All continuous scalar variables now use $\mathcal{G}_M = \{0, 1/M, \dots, 1\}$. Put $d = d_A$ for the joint problem and $d = \max_h(A_{i(h)} - 1)$ for the response-only problem. The finite constraints are

$$\begin{aligned} |v_i(s) - \mathcal{B}_i^{\pi} v_i(s)| &\le \beta_M, & \mathcal{B}_i^{a_i, \pi_{-i}} u_i(s) - u_i(s) &\le \beta_M, \\ |v_h(s) - \mathcal{B}_h^{\sigma_h} v_h(s)| &\le \beta_M, & \mathcal{B}_h^{a} u_h(s) - u_h(s) &\le \beta_M, \end{aligned} \tag{31}$$

where $\beta_M = (d + 1)/M$. Only the applicable row is present for each head. Selectors and simplexes are not relaxed. Minimize the maximum of the local costs $[u_g(s) - v_g(s)]_+$. The optimum is c_M. Every finite variable has domain size at most $D_M = \max\{K, M + 1\}$.

E.2 Rounding a continuous optimum

Lemma E.1 (Simplex and product rounding)**.** *Every distribution on A actions has a denominator-M distribution at TV distance at most $(A-1)/M$. Rounding a product profile gives joint-action TV distance at most d_A/M at every state.*

Proof. For $a < A$, replace p_a by $\lfloor Mp_a \rfloor / M$, and assign the remaining mass to action A. Write $e_a = p_a - \lfloor Mp_a \rfloor / M$, with $0 \le e_a < 1/M$. The final action probability increases by $\sum_{a<A} e_a$. The total ℓ_1 change is therefore $2\sum_{a<A} e_a$, and TV is half that change. Nonnegativity and unit sum are preserved.

For distributions $p_1, \dots, p_n$ and their rounded counterparts $\widetilde{p}_1, \dots, \widetilde{p}_n$, replace one factor at a time. Tensoring a signed measure with a probability distribution preserves its total variation norm. The triangle inequality then bounds TV of the products by $\sum_i \mathrm{TV}(p_i, \widetilde{p}_i) \le d_A/M$. The same calculation over opponents alone is no larger. A declared response task rounds only its focal distribution, which gives the response-only value of d. □

Choose an attained optimum of the continuous problem. Attainment follows from finite partitions, compact policy simplexes, and continuity of discounted values; Appendix C provides the details. Retain its labels. Round code probabilities by Lemma E.1 and copy them to their local occurrences, preserving every selector. Round each normalized witness to a nearest point of $\mathcal{G}_M$. The coordinate error is at most $1/(2M)$, including at the box boundaries.

Fix one state and one Bellman constraint. Changing only its witness changes its one-step successor expectation by at most $\gamma/(2M)$; changing the current-state subtraction contributes at most $1/(2M)$. For the remaining actor perturbation, the one-step target $(1-\gamma)r + \gamma P\widetilde{v}$ is in $[0,1]$. Its expectation changes by at most the TV distance of the actor distributions, hence by at most d/M. Thus the total absolute residual perturbation is bounded by

$$\frac{1+\gamma}{2M} + \frac{d}{M} \le \frac{d+1}{M} = \beta_M. \tag{32}$$

An exact evaluation equality therefore satisfies its relaxed equality. A nonpositive optimal-response residual becomes no larger than β_M. This covers every deployed player, focal action, task, and state simultaneously.

Each difference $u_g(s) - v_g(s)$ increases by at most $1/M$, because each of its two coordinates changes by at most $1/(2M)$. The positive-part function and the maximum are 1-Lipschitz in the sup norm. The rounded assignment is feasible and has objective at most $J_K/H + 1/M$. Consequently

$$c_M \le J_K/H + 1/M, \qquad [Hc_M - H/M]_+ \le J_K. \tag{33}$$

The rounded optimum also proves that the finite feasible set is nonempty.

E.3 Decoding a finite assignment

Lemma E.2 (Approximate Bellman witnesses)**.** *Let $\overline{V} = V/H$ and $\overline{W} = W/H$ denote actual normalized selected and optimal-response values. If $v, u \in [0,1]^N$ satisfy*

$$\|v - \mathcal{B}^\sigma v\|_\infty \le b, \qquad u \ge \mathcal{B}^* u - a\mathbf{1},$$

then $\|v - \overline{V}\|_\infty \le Hb$ and $\overline{W} \le u + Ha\mathbf{1}$. Hence

$$W - V \le H(u - v) + H^2(a+b)\mathbf{1}.$$

The same statement applies to a player's deviation against the decoded deployed opponents.

Proof. Policy evaluation has fixed point $\overline{V} = \mathcal{B}^\sigma \overline{V}$. Subtract and insert $\mathcal{B}^\sigma v$:

$$\|v - \overline{V}\|_\infty \le \|v - \mathcal{B}^\sigma v\|_\infty + \|\mathcal{B}^\sigma v - \mathcal{B}^\sigma \overline{V}\|_\infty \le b + \gamma \|v - \overline{V}\|_\infty.$$

Move the last term to the left and divide by $1-\gamma$, giving Hb.

For the upper witness, Bellman operators shift a constant by γ times that constant. Therefore

$$u + Ha\mathbf{1} - \mathcal{B}^*(u + Ha\mathbf{1}) = u - \mathcal{B}^* u + (1-\gamma)Ha\mathbf{1} = u - \mathcal{B}^* u + a\mathbf{1} \ge 0.$$

Thus $u + Ha\mathbf{1}$ is a supersolution. Iterating the monotone contraction from that supersolution bounds its fixed point $\overline{W}$ from above. The supersolution need not itself belong to $[0,1]^N$; boundedness and contraction on all finite vectors suffice. Combining $\overline{W} \le u + Ha\mathbf{1}$ with $\overline{V} \ge v - Hb\mathbf{1}$, subtracting, and multiplying by H gives the final inequality. The response operator allows original-information deviations by Lemma C.1. □

A minimizing grid assignment decodes to valid code-constant actors and heads because selectors and simplexes are exact. Apply Lemma E.2 with $a = b = \beta_M$ to every head. Its difference satisfies $u_g(s) - v_g(s) \le c_M$. Taking all maxima gives

$$F_G(\widehat{f}) \le Hc_M + 2H^2\beta_M.$$

Every unilateral gain and best-response loss is at most H, so clipping the upper endpoint at H is valid. Every decoded system is feasible for the original partition budget, hence $J_K \le F_G(\widehat{f})$. With Equation 33, this proves every deterministic inequality in Theorem 5.1. Subtracting the unclipped endpoints gives

$$\Delta_M = H/M + 2H^2(d+1)/M.$$

Clipping can only shorten the interval.

Optional residual polishing. After computing c_M at the prescribed slack, search smaller nonnegative slacks while requiring objective no larger than c_M. The original lower endpoint remains valid because its proof used the original finite optimum. Every successful tighter assignment gives a new valid upper endpoint by Lemma E.2. More generally, its actual maximum positive optimality residual a_g and evaluation residual b_g can be evaluated and substituted headwise. The implementation tests zero slack and then five bisection steps. Any successful test improves the feasible upper endpoint, while the lower endpoint remains the one proved from the original finite optimum.

E.4 Exact min–max elimination

For the finite rounded problem, encode a violated hard constraint by $+\infty$, a satisfied hard constraint by zero, and each local objective term by its nonnegative cost. Let the finite variables be $x_1, \dots, x_p$, with domains $\mathcal{X}_1, \dots, \mathcal{X}_p$, and let $f_1, \dots, f_q$ be the resulting factors. The rounded optimization problem is exactly

$$c_M = \min_{x_1 \in \mathcal{X}_1, \dots, x_p \in \mathcal{X}_p} \max_{1 \le a \le q} f_a(x_{\mathrm{scope}(a)}). \tag{34}$$

Lemma E.3 (One-variable min–max elimination)**.** *Fix a variable x_v. For a fixed assignment x_R of all remaining variables, let $g(x_R)$ be the maximum of factors not containing x_v, and let $h(x_v, x_R)$ be the maximum of factors containing x_v. Then*

$$\min_{x_v \in \mathcal{X}_v} \max\{g(x_R), h(x_v, x_R)\} = \max\left\{ g(x_R), \min_{x_v \in \mathcal{X}_v} h(x_v, x_R) \right\}. \tag{35}$$

The identity remains valid when a factor value is $+\infty$.

Proof. Set $m(x_R) = \min_{x_v} h(x_v, x_R)$. For every x_v, $h(x_v, x_R) \ge m(x_R)$, hence

$$\max\{g(x_R), h(x_v, x_R)\} \ge \max\{g(x_R), m(x_R)\}.$$

Taking the minimum over x_v proves the "$\ge$" direction. The domain $\mathcal{X}_v$ is finite, so there is an x_v^* attaining $m(x_R)$. Substitution gives

$$\min_{x_v} \max\{g, h(x_v, \cdot)\} \le \max\{g, h(x_v^*, \cdot)\} = \max\{g, m\},$$

which proves equality. The order relation and minimum operation extend to the value $+\infty$ without change. □

Choose an elimination order $v_1, \dots, v_p$. At step r, collect every current factor containing x_{v_r} into one bucket. Let B_r be the union of all other variables in that bucket. Replace the bucket by the message

$$m_r(x_{B_r}) = \min_{x_{v_r}} \max_{f \in \mathcal{B}_r} f(x_{\mathrm{scope}(f)}). \tag{36}$$

Store, for every assignment to x_{B_r}, one minimizer $x_{v_r}^*(x_{B_r})$.

Lemma E.4 (Exactness of elimination and backtracking)**.** *After eliminating $v_1, \dots, v_r$, the maximum of the remaining original factors and generated messages has, for every fixed assignment of the uneliminated variables, the same minimum over the eliminated variables as the original factor maximum. After all variables are eliminated, the final scalar equals c_M. Reverse-order substitution of the stored minimizers reconstructs a global minimizer of* (34)*.*

Proof. We prove the first statement by induction on r.

For $r = 0$ it is the definition of the original problem. Assume it holds after $r - 1$ steps and consider x_{v_r}. For a fixed assignment of variables that remain after eliminating x_{v_r}, split the current factor maximum into (i) the maximum g of factors not containing x_{v_r} and (ii) the maximum h of the bucket factors that do contain it. Lemma E.3 shows that minimizing over x_{v_r} is exactly equivalent to replacing h by the message $m_r = \min_{x_{v_r}} h$. Hence the induction invariant is preserved.

After p steps no variables remain, so the invariant states that the remaining scalar equals the minimum of the original factor maximum, namely c_M.

For reconstruction, at the final step choose the stored minimizer for the empty separator. Suppose variables $v_{r+1}, \dots, v_p$ have already been reconstructed. Their values determine the separator assignment x_{B_r}, because every variable in B_r occurs later in the elimination order. Set $x_{v_r} = x^*_{v_r}(x_{B_r})$. By definition of the stored minimizer, this choice attains the message value used in the forward elimination. Induction backward through the order reconstructs an assignment attaining the final scalar c_M. □

E.5 Treewidth, runtime, and bit complexity

The primal graph of the finite factorization has one vertex per finite variable and an edge between two variables whenever they occur together in a factor. Eliminating a vertex connects all of its uneliminated neighbors; the number of those neighbors at the time of elimination is the induced width of that step. The width of an order is the maximum such number. The treewidth relevant to the algorithm is the minimum induced width over elimination orders on this primal graph.

Consider a supplied elimination order of width w. The bucket of an eliminated variable contains that variable and at most w uneliminated neighbors, so its union scope has at most $w + 1$ variables. Every finite variable has at most

$$D_M = \max\{K, M + 1\}$$

values. Therefore a complete enumeration of one bucket table requires at most D_M^{w+1} assignments. Evaluating the original local factors at an assignment takes polynomial time in the explicit game/task input, and taking a maximum and a minimum adds only polynomial overhead. There are polynomially many original variables and factors. Assigning every original factor to the earliest variable in its scope prevents repeated evaluation across unrelated buckets. Thus the complete forward elimination and backtracking cost is

$$\mathrm{poly}(\mathcal{I}, \log M) D_M^{w+1}. \tag{37}$$

The message stored after eliminating a variable has at most D_M^w entries; storing all messages is bounded by the same asymptotic expression.

For rational input, every grid probability and normalized witness is rational with denominator dividing a power of M times the input denominators. Each Bellman factor uses a number of additions and multiplications polynomial in the explicitly listed actions and successors. Hence each factor value has bit length polynomial in $\mathcal{I}$ and $\log M$. Multiplying a factor comparison by a positive common denominator converts it to an exact integer comparison. The min and max operations do not increase the bit length beyond the largest represented input value and the accumulated rational-arithmetic overhead. Consequently (37) is also a bit-complexity bound up to a polynomial factor.

A tree decomposition of width w gives an elimination order of width at most w by repeatedly eliminating a vertex when it is forgotten from a leaf bag. Conversely, an elimination order records its achieved width through the filled neighbor sets. The algorithm can therefore use either a supplied tree decomposition or a directly verified elimination order.

E.6 DATA-TO-POPULATION TRANSFER

Assume an event uniformly bounds all stationary joint-policy values between G and an empirical game $\widehat{G}$ by e. Appendix G proves in detail that

$$|F_G(f) - F_{\widehat{G}}(f)| \le 2e, \qquad |J_K(G, \mathcal{H}) - J_K(\widehat{G}, \mathcal{H})| \le 2e \tag{38}$$

for every data-selected system and stationary task family. Hence an empirical lower–upper interval $[L_{\text{emp}}, U_{\text{emp}}]$ transfers to $[L_{\text{emp}} - 2e, U_{\text{emp}} + 2e]$, clipped below at zero. The statistical event and the factor-graph computation control separate sources of uncertainty.

E.7 COMPLETION OF THE PROOF OF THEOREM 5.1

We now assemble the preceding lemmas.

Step 1: a valid lower endpoint. Equation 33 gives

$$L_M := \left[Hc_M - \frac{H}{M} \right]_+ \le J_K(G, \mathcal{H}).$$

This follows by rounding an attained continuous optimum outward into the finite problem.

Step 2: a feasible upper endpoint. Lemma E.4 computes a minimizing finite assignment exactly. Decode its code labels and mixed actions. Selectors and simplex constraints are exact, so the decoded policies are feasible for the original K-code system. Lemma E.2 with $a = b = \beta_M = (d+1)/M$ gives

$$U_M := F_G(\widehat{f}) \le Hc_M + \frac{2H^2(d+1)}{M}.$$

Since $\widehat{f}$ is feasible, $J_K(G, \mathcal{H}) \le U_M$.

Step 3: deterministic interval width. Ignoring the lower clipping, which can only shorten the interval,

$$U_M - L_M \le \frac{H}{M} + \frac{2H^2(d+1)}{M}. \tag{39}$$

For the joint problem, the free local probability coordinates of the deployment actors form a clique in the induced primal graph, so $d = d_A \le w + 1$. Also $H \ge 1$. Therefore

$$H + 2H^2(d+1) \le H^2 + 2H^2(w+2) = H^2(2w+5).$$

It is enough to choose

$$M \ge \frac{H^2(2w+5)}{\varepsilon}. \tag{40}$$

Then (39) is at most ε and

$$L_M \le J_K(G, \mathcal{H}) \le F_G(\widehat{f}) \le L_M + \varepsilon.$$

Step 4: domain size. For any nontrivial head, the selector subgraph contains $K_{N,K}$ between local and abstract probability variables, so $K \le w$, as proved below. From (40) and $0 < \varepsilon \le H$, $M + 1$ dominates K after increasing the universal constant if needed. Thus

$$D_M \le 2 + \frac{H^2(2w+5)}{\varepsilon}.$$

Substituting into (37) yields the runtime in Equation 9.

Step 5: arbitrary action-dependent transitions. The construction of every finite Bellman factor uses the transition probabilities exactly as listed in the input. No factorization, action independence, or low-rank condition is used in the rounding, elimination, or decoding arguments. This completes the proof of Theorem 5.1.

E.8 WHY THE WIDTH PARAMETER IS STRUCTURAL

E.8.1 PARTITION BUDGET AND PROBABILITY DIMENSION

Assume some head has at least two actions. Pick one of its free local probabilities p_s at each state and its matching global probabilities x_k for each code. Every selector contains p_s and x_k, so the primal graph contains all edges of $K_{N,K}$. With $K \le N$, this graph has a $K+1$ clique minor: retain K left vertices and K right vertices, contract the first $K-1$ matched pairs, and leave the last left and right vertices uncontracted. Every pair of resulting vertices has an edge. A clique on $K+1$ vertices has treewidth K, and eliminating/contracting graph edges cannot increase minimum treewidth. Hence every order of the original graph has width at least K.

The deployment-evaluation factor at any state includes the first $A_i - 1$ probability coordinates of every active player. Those d_A coordinates form a clique. Thus $d_A \le w + 1$ (the additional witness often gives a stronger bound). The response-only analogue uses the largest focal alphabet. These claims concern the stated canonical graph before semantic cancellation; if a different exact reformulation is used, its actual domain sizes and achieved width must be reported instead.

For $0 < \varepsilon \le H$, choosing $M = \lceil H^2(2w+5)/\varepsilon \rceil$ dominates both K and the required approximation denominator. Indeed,

$$H + 2H^2(d_A + 1) \le H^2 + 2H^2(w+2) = H^2(2w+5),$$

using $H \ge 1$. Also $H^2(2w+5)/\varepsilon \ge H(2w+5) \ge K$. Substituting into the domain-size bound proves Equation 9. Factors polynomial in $\log M$ can be absorbed into the parameter-dependent multiplier. The resulting parameter dependence is fixed-parameter in the joint tuple $(w, H, 1/\varepsilon)$; the exponent of the inverse-accuracy term scales with w. If every head has one action, all losses vanish and one code suffices directly.

E.9 AN UNBOUNDED-STATE FAMILY WITH BOUNDED PRIMAL TREEWIDTH

Consider a path-structured family whose induced primal graph has bounded treewidth for arbitrarily large state spaces. Let states lie on a path, with every action-dependent transition supported on the current state and its immediate neighbors. Rewards and the transition probabilities on this support may be arbitrary. Fix the player action sizes, partition budget, and number of task heads. Write

$$D_{\mathrm{head}} = \sum_i (A_i - 1) + \sum_h (A_{i(h)} - 1), \qquad b = 1 + D_{\mathrm{head}} + 2(n + |\mathcal{H}|).$$

There are KD_{head} global code coordinates and b variables per state. A path decomposition whose bag at location s contains all global code coordinates and the blocks for $s-1, s, s+1$ has size at most $KD_{\mathrm{head}} + 3b$. Every selector and simplex lies in one such bag. Every Bellman factor uses local probabilities and witness coordinates of these three states, so it also lies in one bag. Every local gap lies there. A state block occurs in three consecutive bags, while global coordinates occur in all bags, proving the running-intersection condition. Thus

$$\mathrm{tw}(\mathrm{primal}(\mathcal{G}_{\mathrm{B}})) \le KD_{\mathrm{head}} + 3b - 1,$$

independently of N. Transitions may point both directions, have action-dependent probabilities, and contain self-loops; they need not be acyclic, exogenous, or low rank. Strict diagonal dominance supplies full-rank examples on the same support.

Conversely, a dense successor support makes one canonical Bellman factor contain N witness coordinates, forcing width at least $N-1$. More importantly, the response-coloring reduction in Proposition D.3 has only self-loop transitions—a physical graph with no inter-state edges—but its shared-code response problem is NP-complete. Thus physical transition sparsity and induced primal-graph treewidth measure different computational structure. Adding a declared head also adds variables and factors; along any retained ordering its induced coupling can grow. The task count therefore contributes to both input size and factorization width.

E.10 BINARY RESPONSE-WIDTH SPECIALIZATION

We prove Equation 41 and distinguish the two widths. Assume one binary focal alphabet and a finite stationary task family. Compute each W_h by its finite discounted MDP linear program. This step is

polynomial in the explicit rational tables and bit length. Compute the two Bellman deficits at every state and record ties exactly.

Appendix D.2 proves that, for a code cell, the optimal one-step mixed deficit is the largest harmonic conflict weight $g_h(s)g_h(t)/(g_h(s)+g_h(t))$ over opposite singleton optimal-action sets. Denote by G_{C} the union of positive-weight edges. At fixed labels, $b_{\mathcal{H}}(\phi)$ is exactly the maximum weight of a monochromatic edge. Give each edge factor value zero if its endpoint labels differ and its weight otherwise. Min–max elimination on a width-c order of G_{C} therefore computes

$$b_K = \min_{\phi:\,|\phi|\le K} \max_{\phi(s)=\phi(t)} \omega(s,t)$$

exactly in $\mathrm{poly}(\mathcal{I})K^{c+1}$ time. The factor weights can be stored by their ordered ranks, so no approximation of rational edge weights is needed.

From the returned labels, construct a mixed response for each cell and task. If p is its probability of action zero, a state uniquely preferring zero requires $p \ge 1 - b_K/g_h(s)$; a state preferring one requires $p \le b_K/g_h(s)$. Ties impose no condition. Intersect these intervals with $[0,1]$. Every lower endpoint is at most every upper endpoint precisely because all same-cell opposite pairs have weight at most b_K. Thus the intersection is nonempty; any point in it defines the cell head. Its statewise Bellman deficit is at most b_K. The exact identity

$$W_h - V_h^{\sigma} = (I - \gamma P_h^{\sigma})^{-1}(W_h - T_h^{\sigma} W_h)$$

has a nonnegative resolvent with row sums H, even when transitions depend arbitrarily on actions. It follows that $S_{\mathcal{H}}(\widehat{\phi}) \le H b_K$.

For any code policy, the same resolvent begins with the identity, so its loss is at least its one-step deficit. Minimizing over heads and encoders gives $b_K \le D_K$. The decoded construction is a feasible encoder/head system, so $D_K \le S_{\mathcal{H}}(\widehat{\phi})$. Therefore

$$b_K \le D_K(G,\mathcal{H}) \le S_{\mathcal{H}}(\widehat{\phi}) \le H b_K. \tag{41}$$

The chain also holds when $b_K = 0$. The one-step proxy is used to construct and bound a response; it is not substituted for the nonzero best-response loss.

If $K \ge c+1$, reverse the elimination order and greedily color each vertex differently from its at most c later neighbors in the filled graph. This gives a proper coloring of the original conflict graph and zero loss. Otherwise $K \le c$, and the preceding runtime is bounded by $\mathrm{poly}(\mathcal{I})(c+1)^{c+1}$. This is fixed-parameter tractable in c for the H-approximation and exact zero-feasibility problem, with no inverse-accuracy parameter in the exponent. For $c = 0$, there are no conflicts and one color suffices. General $A > 2$ lacks the pairwise intersection property: the three sets $\{1,2\},\{2,3\},\{1,3\}$ intersect pairwise but not jointly. The binary graph theorem is not applied to that case.

Exact zero-loss preprocessing. States with identical vectors of optimal-action sets across tasks may be represented by one type in the zero-loss problem. A feasible full-state partition induces at least one used color for each type. Assign every state of a type to a color used by any one representative: that type permits exactly the same actions as the representative in every task, so no common-action intersection becomes empty. Conversely, a type partition lifts to the states. The minimum zero-loss code count is unchanged. This reduces some large tactical examples before coloring, but need not preserve nonzero losses, deployment compatibility, or full primal treewidth induced by the factor graph $\mathcal{G}_{\mathrm{B}}$. Those targets continue to use Theorem 5.1.

F Mixed-Integer Outer Relaxation for High-Width Games

This appendix gives an alternative to the finite-grid elimination algorithm for the same joint objective J_K. The outer relaxation retains continuous policy probabilities and uses a binary expansion with a bounded remainder. All reported objective endpoints are in unnormalized return units.

Theorem F.1 (Joint optimization with an observable optimality interval)**.** *For a rational finite discounted game, finite rational response-task family, and partition budget K, define*

$$d_A = \sum_i (|\mathcal{A}_i| - 1), \qquad d_* = \max_i \{2d_A - (|\mathcal{A}_i| - 1)\}, \qquad \Delta_M^{\mathrm{out}} = \frac{H^2 d_*}{2M}. \tag{42}$$

The joint Bellman program admits a mixed-integer linear outer relaxation indexed by $M \geq 1$. If ℓ is a valid global lower bound and an integer-feasible incumbent has objective t_M, its decoded policy system $\widehat{f}$ satisfies

$$\ell \leq J_K(\mathcal{G}, \mathcal{Q}) \leq u := F_{\mathcal{G}}(\widehat{f}) \leq t_M + \Delta_M^{\text{out}}. \tag{43}$$

For response-only synthesis, replace d_* by $\max_q(|\mathcal{A}_{i(q)}| - 1)$. The formulation size is polynomial in the explicit input and $\log M$; solving it can require exponential time.

F.1 Shared partition and simplex-preserving probability expansion

Use indicators $y_{sk} \in \{0,1\}$ with $\sum_{k=1}^{K} y_{sk} = 1$, and code probabilities $x_{k,g,a} \geq 0$ with $\sum_a x_{k,g,a} = 1$ for every deployment or response head g. Conditional equalities copy the selected code probabilities to each state. Empty cells are permitted. For each code and head, suppressing those indices, expand the first $A - 1$ probabilities as

$$x_a = \sum_{j=0}^{L-1} \frac{2^j}{M} b_{aj} + e_a, \quad b_{aj} \in \{0,1\}, \quad 0 \leq e_a \leq \delta_M := 1/M, \quad L = 1 + \lfloor \log_2 M \rfloor, \tag{44}$$

with $x_A = 1 - \sum_{a<A} x_a \geq 0$. Every simplex point is represented by the binary expansion of $\lfloor M x_a \rfloor$ and the remainder $x_a - \lfloor M x_a \rfloor / M$. State copies of these variables agree within each cell.

All value witnesses in this construction are normalized to $[0,1]$. The normalized Bellman operator is $\mathcal{B}_i^{\mathbf{a}} v = (1-\gamma) r_i^{\mathbf{a}} + \gamma P^{\mathbf{a}} v$. Multiplying the final objective by H restores return units.

F.2 Exact binary products and remainder envelopes

For $b \in \{0,1\}$ and $v \in [0,1]$, the product $z = bv$ is represented exactly by

$$0 \leq z \leq b, \qquad z \leq v, \qquad z \geq v + b - 1.$$

For $e \in [0, \delta_M]$, the McCormick envelope for $w = ev$ is

$$w \geq 0, \quad w \geq e + \delta_M v - \delta_M, \quad w \leq e, \quad w \leq \delta_M v. \tag{45}$$

Every exact product lies in this envelope. Rescaling $a = e/\delta_M$ gives

$$|w - ev| \leq \delta_M / 4.$$

Indeed, the upper error is $\delta_M[\min(a, v) - av] \leq \delta_M/4$; the lower error is $\delta_M[av - \max(0, a + v - 1)] \leq \delta_M/4$.

Represent $x_a v$ for $a < A$ by the exact digit products plus the remainder envelope. Define the last-action product by

$$\widehat{x_A v} = v - \sum_{a<A} \widehat{x_a v}.$$

This preserves the simplex identity. Its error is at most $(A-1)/(4M)$, while each other action product has error at most $1/(4M)$. Therefore, for any action-value table $(v_a) \in [0,1]^A$,

$$\left| \sum_a \widehat{x_a v_a} - \sum_a x_a v_a \right| \leq c_A := \frac{A-1}{2M}. \tag{46}$$

Intermediate averages are constrained to $[0,1]$, as every exact average lies in that interval.

F.3 Nested independent-action averages

For each state and joint action, the continuation target

$$Q_i(s, \mathbf{a}) = (1-\gamma) r_i(s, \mathbf{a}) + \gamma \sum_{s'} P(s' \mid s, \mathbf{a}) v_i(s')$$

is linear in the witnesses and lies in $[0,1]$. Average this table over the players one at a time using (46). Averaging an existing error with a probability vector does not increase its supremum norm; each new player adds at most $c_{|\mathcal{A}_i|}$. Thus the deployment evaluation incurs error at most

$$c_{\text{all}} = \frac{d_A}{2M},$$

whereas player i's optimal-response inequalities incur at most

$$c_{-i} = \frac{d_A - (|\mathcal{A}_i| - 1)}{2M}.$$

The focal action is held fixed in the latter averages. For a declared response task, opponent probabilities are fixed: the optimal-response inequalities are linear and exact, and only response evaluation incurs $c_{|\mathcal{A}_{i(q)}|}$.

F.4 Proof of the optimality interval

Proof of Theorem F.1. Every continuous partition-policy system embeds in the lift: use (44), exact products, and exact normalized value witnesses. Hence the outer optimum is no greater than J_K, and every global lower bound ℓ is valid.

An integer-feasible incumbent decodes to a shared partition and valid continuous policies. Let its normalized objective be $t = t_M/H$. For deployment player i, the relaxed witnesses satisfy

$$\|v_i - \mathcal{B}_i^{\pi} v_i\|_\infty \le c_{\mathrm{all}}, \qquad u_i \ge \mathcal{B}_i^{*,\pi_{-i}} u_i - c_{-i}\mathbf{1}, \qquad u_i - v_i \le t\mathbf{1}.$$

Lemma E.2 then yields

$$W_i^{\pi_{-i}} - V_i^{\pi} \le \left[t_M + H^2(c_{\mathrm{all}} + c_{-i})\right] \mathbf{1}.$$

For response task q, the analogous bound is $t_M + H^2 c_{|\mathcal{A}_{i(q)}|}$. Taking the maximum over deployment and response heads gives $F_{\mathcal{G}}(\widehat{f}) \le t_M + \Delta_M^{\mathrm{out}}$. Feasibility gives $J_K \le F_{\mathcal{G}}(\widehat{f})$, proving (43).

The number of assignment, digit, product, and intermediate-average variables is polynomial in the explicit game and task tables and $\log M$. For rational inputs, fixing all binary variables leaves a rational linear program. Enumerating binary assignments gives a finite exact procedure, with exponential worst-case cost. □

F.5 Convergence and statistical transfer

Choose M so that $\Delta_M^{\mathrm{out}} \le \xi/2$ and solve to branch-and-bound gap $t_M - \ell \le \xi/2$. Then the decoded upper endpoint satisfies $u - \ell \le \xi$. Keeping the largest valid lower endpoint and smallest feasible upper endpoint across successive meshes gives nested certified intervals.

If the lift uses an empirical model and a uniform value-error event holds with radius e, Lemma G.1 gives

$$[\ell - 2e]_+ \le J_K(\mathcal{G}, \mathcal{Q}) \le u + 2e, \qquad F_{\mathcal{G}}(\widehat{f}) - J_K(\mathcal{G}, \mathcal{Q}) \le u - \ell + 4e.$$

Appendix G specifies sample and solver allocations for this interval.

G Statistical Guarantees and Offline Sample Complexity

This appendix derives the statistical population guarantees when the state partition ϕ, deployment policy π, and task-specific response heads $\{\sigma_q\}_{q\in\mathcal{Q}}$ are computed from an empirical game model $\widehat{\mathcal{G}}$ estimated from finite data. The central theoretical mechanism is a uniform concentration event that controls the value error across all stationary joint policies simultaneously; once this event holds, all downstream abstraction and optimization bounds transfer directly to the true environment $\mathcal{G}$.

G.1 Uniform Model Error Transfers to Best-Response Loss and Nash Gap

Lemma G.1 (Uniform Value Error Transfers the Joint Objective)**.** *Suppose that, on an event $\mathcal{E}$, for every stationary joint policy $\tau \in \Pi$ and every player $i \in \mathcal{N}$,*

$$\|V_{\mathcal{G},i}^{\tau} - V_{\widehat{\mathcal{G}},i}^{\tau}\|_\infty \le e.$$

Then, on $\mathcal{E}$, for every benchmark opponent profile $q \in \mathcal{Q}$, every stationary focal policy $p \in \Pi_i$, every state partition ϕ, and every ϕ-abstract response $\sigma \in \Pi_{\phi,i}$:

$$\|W^q_{\mathcal{G},i} - W^q_{\widehat{\mathcal{G}},i}\|_\infty \le e, \tag{47}$$

$$\left|D_{\mathcal{G},q}(p) - D_{\widehat{\mathcal{G}},q}(p)\right| \le 2e, \tag{48}$$

$$\left|\|W^q_{\mathcal{G},i} - V^{\sigma,q}_{\mathcal{G},i}\|_\infty - \|W^q_{\widehat{\mathcal{G}},i} - V^{\sigma,q}_{\widehat{\mathcal{G}},i}\|_\infty\right| \le 2e. \tag{49}$$

Consequently, for every candidate tuple $f = (\phi, \pi, \{\sigma_q\})$,

$$|F_{\mathcal{G}}(f) - F_{\widehat{\mathcal{G}}}(f)| \le 2e, \qquad |J_K(\mathcal{G}, \mathcal{Q}) - J_K(\widehat{\mathcal{G}}, \mathcal{Q})| \le 2e.$$

All statements hold simultaneously for data-dependent state abstractions, deployment policies, response heads, and benchmark opponent sets evaluated on the event $\mathcal{E}$.

Proof. Fix a focal player i and stationary opponent profile $q \in \mathcal{Q}$.

Step 1: Transfer the unrestricted best-response value. In a finite discounted MDP, a stationary deterministic policy attains the optimal value. Therefore, for every state $s \in \mathcal{S}$:

$$W^q_{\mathcal{G},i}(s) = \max_{\sigma \in \Pi_i} V^{\sigma,q}_{\mathcal{G},i}(s), \qquad W^q_{\widehat{\mathcal{G}},i}(s) = \max_{\sigma \in \Pi_i} V^{\sigma,q}_{\widehat{\mathcal{G}},i}(s).$$

For any two real-valued families $\{a_\sigma\}$ and $\{b_\sigma\}$, $|\sup_\sigma a_\sigma - \sup_\sigma b_\sigma| \le \sup_\sigma |a_\sigma - b_\sigma|$. Applying this inequality coordinatewise and conditioning on the event $\mathcal{E}$ yields:

$$|W^q_{\mathcal{G},i}(s) - W^q_{\widehat{\mathcal{G}},i}(s)| \le \max_{\sigma \in \Pi_i} |V^{\sigma,q}_{\mathcal{G},i}(s) - V^{\sigma,q}_{\widehat{\mathcal{G}},i}(s)| \le e.$$

Taking the supremum over $s \in \mathcal{S}$ proves (47).

Step 2: Transfer unilateral deviation gain. For a stationary focal policy $p \in \Pi_i$, $D_{\mathcal{G},q}(p) = \|W^q_{\mathcal{G},i} - V^{p,q}_{\mathcal{G},i}\|_\infty$ because the discrepancy is coordinatewise non-negative by Bellman optimality. Applying the reverse triangle inequality yields:

$$\begin{aligned} |D_{\mathcal{G},q}(p) - D_{\widehat{\mathcal{G}},q}(p)| &\le \|W^q_{\mathcal{G},i} - W^q_{\widehat{\mathcal{G}},i}\|_\infty + \|V^{p,q}_{\mathcal{G},i} - V^{p,q}_{\widehat{\mathcal{G}},i}\|_\infty \\ &\le e + e = 2e, \end{aligned}$$

which establishes (48).

Step 3: Transfer the best-response loss. The identical reverse-triangle bound applied with p replaced by an arbitrary ϕ-abstract policy $\sigma \in \Pi_{\phi,i}$ yields (49). This transfer holds regardless of whether σ is optimal in either game model.

Step 4: Transfer joint policy evaluations. The joint system objective $F_{\mathcal{G}}(f) = \max\{\mathrm{Gap}_{\mathcal{G}}(\pi), \max_{q \in \mathcal{Q}} \|W^q_{i(q)} - V^{\sigma_q,q}_{i(q)}\|_\infty\}$ is a maximum of components bounded in Steps 2 and 3. Because each individual evaluation changes by at most $2e$, their maximum changes by at most $2e$: $|F_{\mathcal{G}}(f) - F_{\widehat{\mathcal{G}}}(f)| \le 2e$ for every feasible tuple f.

Step 5: Transfer the joint abstraction optimum. Let $f^*_{\mathcal{G}}$ attain $J_K(\mathcal{G}, \mathcal{Q})$ and $f^*_{\widehat{\mathcal{G}}}$ attain $J_K(\widehat{\mathcal{G}}, \mathcal{Q})$ under partition budget K. Then:

$$J_K(\mathcal{G}, \mathcal{Q}) = F_{\mathcal{G}}(f^*_{\mathcal{G}}) \le F_{\mathcal{G}}(f^*_{\widehat{\mathcal{G}}}) \le F_{\widehat{\mathcal{G}}}(f^*_{\widehat{\mathcal{G}}}) + 2e = J_K(\widehat{\mathcal{G}}, \mathcal{Q}) + 2e.$$

Interchanging $\mathcal{G}$ and $\widehat{\mathcal{G}}$ gives the matching lower bound, confirming $|J_K(\mathcal{G}, \mathcal{Q}) - J_K(\widehat{\mathcal{G}}, \mathcal{Q})| \le 2e$. Because $\mathcal{E}$ holds uniformly across all stationary policies prior to abstraction selection, data-dependent choices inherit the guarantee. □

G.2 Population Optimality Interval

Suppose the empirical optimization solver returns a certified lower bound $\ell \le J_K(\widehat{\mathcal{G}}, \mathcal{Q})$ and a decoded feasible policy system $\widehat{f}$ with independently evaluated empirical objective $u = F_{\widehat{\mathcal{G}}}(\widehat{f})$, satisfying $\ell \le J_K(\widehat{\mathcal{G}}, \mathcal{Q}) \le u$. Applying Lemma G.1 gives:

$$J_K(\mathcal{G}, \mathcal{Q}) \ge J_K(\widehat{\mathcal{G}}, \mathcal{Q}) - 2e \ge \ell - 2e,$$
$$J_K(\mathcal{G}, \mathcal{Q}) \le J_K(\widehat{\mathcal{G}}, \mathcal{Q}) + 2e \le u + 2e,$$

and

$$F_{\mathcal{G}}(\widehat{f}) \le F_{\widehat{\mathcal{G}}}(\widehat{f}) + 2e = u + 2e.$$

Because $J_K \ge 0$, the certified population lower endpoint is $[\ell - 2e]_+$. The suboptimality of the decoded abstraction on the true environment $\mathcal{G}$ satisfies:

$$F_{\mathcal{G}}(\widehat{f}) - J_K(\mathcal{G}, \mathcal{Q}) \le (u + 2e) - (\ell - 2e) = u - \ell + 4e. \tag{50}$$

G.3 Sample Complexity under Unknown Transitions and Rewards

Let $N := |\mathcal{S}|$, $J := |\mathcal{A}| = \prod_{i=1}^{n} |\mathcal{A}_i|$, and $n := |\mathcal{N}|$. At each state-joint-action cell $(s, \mathbf{a})$, suppose the learner collects m_0 conditionally independent transitions and reward observations (R_i, S'), with $R_i \in [0, 1]$.

For any player i, cell $(s, \mathbf{a})$, and target value vector $v \in [0, H]^N$, define the random one-step Bellman target $Y_v := R_i + \gamma v(S')$. Because $R_i \in [0, 1]$ and $v \in [0, H]^N$, $Y_v \in [0, 1 + \gamma H] = [0, H]$. Hoeffding's inequality implies:

$$\Pr\left(\left|(\widehat{r}_i - r_i)(s, \mathbf{a}) + \gamma[(\widehat{P} - P)v](s, \mathbf{a})\right| > \beta\right) \le 2\exp\left(-\frac{2m_0\beta^2}{H^2}\right). \tag{51}$$

The term inside the absolute value is affine in v. The extrema of an affine function over the hypercube $[0, H]^N$ are attained at its 2^N extreme vertices. Union bounding over all nNJ player-cell pairs and all 2^N vertices gives an aggregate failure probability bounded by:

$$2nNJ \cdot 2^N \exp\left(-\frac{2m_0\beta^2}{H^2}\right).$$

Setting this failure budget to $\alpha \in (0, 1)$ yields:

$$\beta = H\sqrt{\frac{N\log 2 + \log(2nNJ/\alpha)}{2m_0}} = H\sqrt{\frac{C_\alpha}{2m_0}}, \qquad C_\alpha := N\log 2 + \log(2nNJ/\alpha). \tag{52}$$

Thus, with probability at least $1 - \alpha$:

$$\sup_{i \in \mathcal{N}, s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} \sup_{v \in [0,H]^N} \left|(\widehat{r}_i - r_i)(s, \mathbf{a}) + \gamma[(\widehat{P} - P)v](s, \mathbf{a})\right| \le \beta. \tag{53}$$

For any stationary joint policy τ, averaging (53) over $\tau(\cdot \mid s)$ implies $\|\mathcal{T}^\tau_{\mathcal{G}} v - \mathcal{T}^\tau_{\widehat{\mathcal{G}}} v\|_\infty \le \beta$. Using the fixed-point identities $V^\tau_{\mathcal{G}} = \mathcal{T}^\tau_{\mathcal{G}} V^\tau_{\mathcal{G}}$ and $V^\tau_{\widehat{\mathcal{G}}} = \mathcal{T}^\tau_{\widehat{\mathcal{G}}} V^\tau_{\widehat{\mathcal{G}}}$:

$$\begin{aligned}\|V^\tau_{\mathcal{G}} - V^\tau_{\widehat{\mathcal{G}}}\|_\infty &\le \|\mathcal{T}^\tau_{\mathcal{G}} V^\tau_{\mathcal{G}} - \mathcal{T}^\tau_{\mathcal{G}} V^\tau_{\widehat{\mathcal{G}}}\|_\infty + \|\mathcal{T}^\tau_{\mathcal{G}} V^\tau_{\widehat{\mathcal{G}}} - \mathcal{T}^\tau_{\widehat{\mathcal{G}}} V^\tau_{\widehat{\mathcal{G}}}\|_\infty \\ &\le \gamma\|V^\tau_{\mathcal{G}} - V^\tau_{\widehat{\mathcal{G}}}\|_\infty + \beta.\end{aligned}$$

Rearranging yields the uniform value deviation bound:

$$\|V^\tau_{\mathcal{G}} - V^\tau_{\widehat{\mathcal{G}}}\|_\infty \le \frac{\beta}{1 - \gamma} = H\beta. \tag{54}$$

Because value functions are bounded in $[0, H]$, the uniform concentration event $\mathcal{E}$ in Lemma G.1 holds with:

$$e = \min\left\{H,\ H^2\sqrt{\frac{C_\alpha}{2m_0}}\right\}. \tag{55}$$

G.4 Offline Data with Random State-Action Visitation Counts

Suppose L transitions are sampled i.i.d. from a logging distribution $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ satisfying $\mu(s, \mathbf{a}) \geq \nu_{\min} > 0$ for all cells. Let N_x denote the realized visitation count of cell $x = (s, \mathbf{a})$. By the multiplicative Chernoff bound:

$$\Pr\left(N_x < \frac{L\mu(x)}{2}\right) \leq \exp\left(-\frac{L\mu(x)}{8}\right) \leq \exp\left(-\frac{L\nu_{\min}}{8}\right).$$

Taking a union bound over all $D = NJ$ state-action cells reveals that setting:

$$L \geq \frac{8}{\nu_{\min}} \log\left(\frac{2D}{\alpha}\right) \tag{56}$$

ensures $N_x \geq \frac{L\nu_{\min}}{2}$ uniformly across all cells with probability at least $1 - \alpha/2$. Conditioning on this coverage event, the empirical estimation error bounds of Equation (51) apply with $m_0 = \min_x N_x$ under failure probability $\alpha/2$, guaranteeing that the aggregate population bound holds with probability at least $1 - \alpha$.

For sequential online data collection across rounds $r = 1, 2, \ldots$, allocating failure budgets $\alpha_r = \frac{\alpha}{r(r+1)}$ ensures $\sum_{r=1}^{\infty} \alpha_r = \alpha$ by telescoping summation. The population confidence interval holds uniformly across all adaptation rounds simultaneously with probability at least $1 - \alpha$.

G.5 Known Transitions and Variance-Adaptive Reward Confidence

When transition dynamics are known a priori and only reward functions must be estimated, sample requirements can be tightened using empirical Bernstein bounds (Maurer & Pontil, 2009). Let $D_R = nNJ$ count player–state–joint-action entries, and let x index these reward entries. For $N_x \geq 2$, define:

$$\beta_H(x) := \sqrt{\frac{\log(4D_R/\alpha)}{2N_x}}, \qquad \beta_{EB}(x) := \sqrt{\frac{2\widehat{v}_x \log(8D_R/\alpha)}{N_x}} + \frac{7\log(8D_R/\alpha)}{3(N_x - 1)},$$

where $\widehat{v}_x$ is the sample variance of observed rewards at cell x. Allocating failure probability $\alpha/2$ to each bound, the confidence radius satisfies:

$$\beta_{\text{safe}} := \max_x \min\left\{\beta_H(x),\ \beta_{EB}(x)\right\} \tag{57}$$

with probability at least $1 - \alpha$. The resulting policy value transfer bound holds with $e = H\beta_{\text{safe}}$.

G.6 Explicit Resource Allocation

For the mixed-integer outer relaxation of Appendix F, the population interval width decomposes into three additive sources of error: model estimation error $4e$, discretization envelope $\Delta_M^{\text{out}} \leq \frac{H^2 d_*}{2M}$, and branch-and-bound optimization gap $t_M - \ell$. Here $d_* := \max_i\{2d_A - (|\mathcal{A}_i| - 1)\}$, and t_M, ℓ are in unnormalized return units. To guarantee that each source contributes at most $\xi/3$ toward a target tolerance $\xi > 0$, it suffices to allocate:

$$4H^2\sqrt{\frac{C_\alpha}{2m_0}} \leq \frac{\xi}{3}, \qquad \frac{H^2 d_*}{2M} \leq \frac{\xi}{3}, \qquad t_M - \ell \leq \frac{\xi}{3}.$$

Solving for the required sample count per cell m_0 and grid denominator M yields:

$$m_0 \geq \frac{72H^4 C_\alpha}{\xi^2}, \qquad M \geq \frac{3H^2 d_*}{2\xi}. \tag{58}$$

G.7 Selecting the State Partition Budget

Define the minimal partition cardinality required to achieve tolerance ε as:

$$K^*(\varepsilon) := \min\{K \in \mathbb{N} \mid J_K(\mathcal{G}, \mathcal{Q}) \leq \varepsilon\}.$$

Suppose optimization over candidate budgets yields certified intervals $[L_K, U_K]$ of width at most $\xi < \varepsilon$, and let $\widehat{K}$ denote the smallest budget satisfying $U_{\widehat{K}} \le \varepsilon$. Since $J_{\widehat{K}} \le U_{\widehat{K}} \le \varepsilon$, feasibility implies $K^*(\varepsilon) \le \widehat{K}$. Conversely, define $K_0 := K^*(\varepsilon - \xi)$. By definition, $J_{K_0} \le \varepsilon - \xi$, which implies $U_{K_0} \le J_{K_0} + \xi \le \varepsilon$. Minimality of $\widehat{K}$ establishes:

$$K^*(\varepsilon) \le \widehat{K} \le K^*(\varepsilon - \xi). \tag{59}$$

Equation (59) converts objective certificate intervals into a rigorous tolerance-band guarantee for representation budget selection.

G.8 Information-Theoretic Coverage Lower Bound

Proposition G.2 (Small Optimal Codebooks Do Not Imply Small Data Requirements)**.** *For every state space cardinality N and budget $T < N$, for every active learning algorithm querying at most T state-action pairs, there exists an N-state, two-action absorbing response game admitting an exact zero-loss single-state abstraction ($K = 1$) on which that algorithm incurs failure probability at least $\frac{1}{2}(1 - T/N)$ when tasked with achieving best-response loss strictly below $\frac{H\Delta}{2}$.*

Proof. Consider an environment where all states self-loop deterministically: $P(s \mid s, a) = 1$. Action 0 yields deterministic reward $1/2$ at all states. Select a hidden state index $j^* \in [N]$ and a hidden sign $\sigma^* \in \{+1, -1\}$ uniformly at random. Action 1 yields reward $1/2$ at all states $s \neq j^*$, and yields reward $1/2 + \sigma^*\Delta$ at the hidden state j^*, with $\Delta \in (0, 1/2)$.

If $\sigma^* = +1$, the constant action-1 policy is optimal at every state and uniquely optimal at the hidden state j^*; if $\sigma^* = -1$, the constant action-0 policy is optimal at every state and uniquely optimal at j^*. In both realizations, an exact zero-loss response policy is supported on a single abstract state ($K = 1$).

Any active sampling algorithm inspecting at most T state-action pairs observes information about σ^* only if it explicitly queries action 1 at the hidden state j^*. Because j^* is chosen uniformly from $[N]$ independently of the algorithm's internal coin flips, the probability that j^* is never queried satisfies:

$$\Pr(j^* \text{ is unqueried}) \ge 1 - \frac{T}{N}.$$

Conditioned on this event, the observation history is statistically identical under $\sigma^* = +1$ and $\sigma^* = -1$. Let $p \in [0, 1]$ denote the returned policy probability of choosing action 1 at state j^*. Because state j^* absorbs, the expected loss under $\sigma^* = +1$ is $H(1 - p)\Delta$, while the loss under $\sigma^* = -1$ is $Hp\Delta$. Achieving worst-case loss strictly below $\frac{H\Delta}{2}$ requires simultaneously $p > 1/2$ and $p < 1/2$, which is impossible. Averaging over both signs yields conditional failure probability at least $1/2$. Multiplying by the probability of missing j^* establishes that the algorithm fails with probability at least $\frac{1}{2}(1 - T/N)$. □

H Population-Based Learning and Opponent Space Representation

This appendix provides the formal game-theoretic interface between strategy-complete state abstraction and population-based learning algorithms, including Policy-Space Response Oracles (PSRO) and league training. We prove the operational error decomposition, establish public-history representations for episode mixtures, and derive metric coverage guarantees for compact opponent spaces.

H.1 Monotonicity under Opponent Population Expansion

In population-based architectures, the benchmark opponent set $\mathcal{Q}_t = \{q_1, \ldots, q_t\} \subset \Pi_{-i}$ expands iteratively as new counter-strategies are discovered. When an opponent profile q_{t+1} is admitted, the response requirements expand to $\mathcal{Q}_{t+1} = \mathcal{Q}_t \cup \{q_{t+1}\}$.

At any fixed partition budget K, the joint abstraction objective satisfies monotonicity:

$$J_K(\mathcal{G}, \mathcal{Q}_t) \le J_K(\mathcal{G}, \mathcal{Q}_{t+1}) \qquad \text{whenever } \mathcal{Q}_t \subseteq \mathcal{Q}_{t+1}. \tag{60}$$

This holds because any state partition ϕ and policy profile feasible for $\mathcal{Q}_{t+1}$ is admissible for $\mathcal{Q}_t$, and the objective takes a maximum over task losses. Consequently, any certified lower bound for $\mathcal{Q}_t$ remains a valid lower bound for $\mathcal{Q}_{t+1}$. A feasible upper certificate for $\mathcal{Q}_{t+1}$ is obtained by optimizing a response head $\sigma_{q_{t+1}} \in \Pi_{\phi,i(q_{t+1})}$ under the shared partition ϕ. If $S_{q_{t+1}}(\phi) = 0$, the existing abstraction is already strategy-complete for the new opponent, and no partition refinement is required.

H.2 Episode-Level Opponent Mixtures and Public-History MDPs

In standard PSRO meta-game formulations, opponents are frequently randomized at the episode level rather than through statewise Markov mixtures. Suppose an opponent index $k \in [L]$ is sampled at the beginning of an episode with prior distribution $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L) \in \Delta([L])$, and controller q^k governs opponent action selection thereafter. The responder observes the public history:

$$h_t = (s_0, \mathbf{a}_0, s_1, \mathbf{a}_1, \ldots, \mathbf{a}_{t-1}, s_t) \in \mathcal{H}_t,$$

but does not directly observe the realized latent index k.

For each opponent index k, define the unnormalized trajectory likelihood weight:

$$\omega_k(h_t) := \lambda_k \prod_{\tau=0}^{t-1} q^k(a_{-i,\tau} \mid h_\tau). \tag{61}$$

By Bayes' rule, the conditional posterior probability of facing opponent k given history h_t is:

$$\Pr(K = k \mid h_t) = \frac{\omega_k(h_t)}{\sum_{j=1}^{L} \omega_j(h_t)}. \tag{62}$$

The environmental transition probabilities and the focal player's action probabilities cancel out of the numerator and denominator because they are conditionally independent of k given history h_t. The opponent's predictive action distribution given public history h_t is therefore:

$$q_{\boldsymbol{\lambda}}(a_{-i} \mid h_t) = \sum_{k=1}^{L} \Pr(K = k \mid h_t) q^k(a_{-i} \mid h_t) = \frac{\sum_{k=1}^{L} \omega_k(h_t) q^k(a_{-i} \mid h_t)}{\sum_{k=1}^{L} \omega_k(h_t)}. \tag{63}$$

Equation (63) demonstrates that the public history h_t serves as a sufficient Markov state for the induced response problem. On a public-history tree truncated at depth T, the tail return beyond horizon T is bounded by $\sum_{t=T}^{\infty} \gamma^t = H\gamma^T$, introducing a truncation error of at most $H\gamma^T$ to the response value.

Analytical Distinction between Episode Mixtures and Statewise Mixtures. Episode-level mixtures must not be conflated with statewise mixed policies $\bar{q}(a_{-i} \mid s) = \sum_k \lambda_k q^k(a_{-i} \mid s)$. Consider two deterministic opponent controllers: controller 0 always plays action 0, while controller 1 always plays action 1, with equal prior weights $\lambda_0 = \lambda_1 = 1/2$. Under an episode-level mixture, observing the first opponent action reveals the opponent identity with posterior probability 1, permitting optimal counteraction with certainty from step 1 onward. In contrast, under statewise averaging, opponent actions remain conditionally independent across time, yielding a persistent predictability deficit. Strategy-complete abstraction explicitly operates on the correct belief-augmented state space for episode mixtures.

H.3 Payoff Equivalence under Exact Policy Compilation

Suppose each stored opponent controller in the empirical population is compiled into a ϕ-abstract policy $\widehat{q}^k \in \Pi_\phi$ satisfying $\widehat{q}^k(\cdot \mid h_t) = q^k(\cdot \mid h_t)$ for all reachable histories h_t. We prove by induction on episode length t that the trajectory distribution is preserved exactly.

At $t = 0$, initial state distributions coincide. Assuming trajectory equivalence holds for all histories of length t, the transition kernel P, reward functions r_i, and compiled policy distributions agree conditionally on h_t. Multiplying the transition probabilities establishes that the joint distribution over

histories h_{t+1} remains identical. Consequently, expected discounted returns under any mixture $\boldsymbol{\lambda}$ are invariant:

$$\mathbb{E}_{\tau\sim(p,q_{\boldsymbol{\lambda}})}\left[\sum_{t=0}^{\infty}\gamma^t r_i(S_t,\mathbf{A}_t)\right]=\mathbb{E}_{\tau\sim(p,\widehat{q}_{\boldsymbol{\lambda}})}\left[\sum_{t=0}^{\infty}\gamma^t r_i(S_t,\mathbf{A}_t)\right],$$

confirming that exact policy compilation preserves the PSRO meta-game payoff matrix.

H.4 Proof of Proposition 6.2 (Population Deviation Guarantee)

We now prove the unified Nash-gap decomposition established in Proposition 6.2.

Proof of Proposition 6.2. For each player, use its benchmark family containing the deployed opponent profile. Let $\pi \in \Pi_\phi$ be the deployed strategy profile returned by the meta-solver, and let $q = \pi_{-i}$ denote the opponents' joint strategy facing player i. For a focal policy $p \in \Pi_{\phi,i}$, define the unrestricted and in-class worst-state deviation gains by

$$D_q(p) := \|W_i^q - V_i^{p,q}\|_\infty, \qquad D_{\phi,q}(p) := \max_{s\in\mathcal{S}}\left\{\sup_{\sigma\in\Pi_{\phi,i}} V_i^{\sigma,q}(s) - V_i^{p,q}(s)\right\}.$$

Choose $\sigma_\eta \in \Pi_{\phi,i}$ with $\|W_i^q - V_i^{\sigma_\eta,q}\|_\infty \le S_q(\phi) + \eta$. For every state s,

$$\begin{aligned} W_i^q(s) - V_i^{p,q}(s) &= \left[W_i^q(s) - V_i^{\sigma_\eta,q}(s)\right] + \left[V_i^{\sigma_\eta,q}(s) - V_i^{p,q}(s)\right] \\ &\le S_q(\phi) + \eta + D_{\phi,q}(p). \end{aligned}$$

Taking the maximum over states and then letting $\eta \downarrow 0$ gives

$$0 \le D_q(p) - D_{\phi,q}(p) \le S_q(\phi) \le \varepsilon_{\mathrm{BR}}.$$

By the stopping condition over the entire abstract deviation class, the in-class deviation is bounded by τ. Accounting for numerical planning residual $\varepsilon_{\mathrm{plan}}$, statistical evaluation error $\varepsilon_{\mathrm{stat}}$, and solver tolerance $\varepsilon_{\mathrm{opt}}$ yields:

$$D_{\phi,q}(p) \le \tau + \varepsilon_{\mathrm{plan}} + \varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{opt}}.$$

Adding $S_q(\phi) \le \varepsilon_{\mathrm{BR}}$ gives:

$$D_q(p) \le \tau + \varepsilon_{\mathrm{BR}} + \varepsilon_{\mathrm{plan}} + \varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{opt}}.$$

Taking the maximum over players gives the Nash-gap bound in Equation (11). □

H.5 Opponent Coverage over Continuous Opponent Classes

A gap computed only over a finite stored population controls that population's deviations; applying the proposition also requires controlling deviations over $\Pi_{\phi,i}$.

Suppose the game requires robustness against a compact class of stationary opponents $\mathcal{Q}_{\mathrm{all}}$. Let $\mathcal{Q} \subset \mathcal{Q}_{\mathrm{all}}$ form a δ-cover under the statewise total variation metric:

$$\forall q \in \mathcal{Q}_{\mathrm{all}},\ \exists \bar{q} \in \mathcal{Q} \quad \text{such that} \quad \sup_{s\in\mathcal{S}} \mathrm{TV}(q(\cdot \mid s), \bar{q}(\cdot \mid s)) \le \delta.$$

For any fixed focal policy $p \in \Pi_i$, the induced stage rewards satisfy $|r_i^{p,q}(s) - r_i^{p,\bar{q}}(s)| \le \delta$ because rewards lie in $[0,1]$. Similarly, the induced transition distributions satisfy $\mathrm{TV}(P_i^{p,q}(\cdot \mid s), P_i^{p,\bar{q}}(\cdot \mid s)) \le \delta$. For any value vector $v \in [0,H]^{|\mathcal{S}|}$, the evaluation operators satisfy:

$$\|\mathcal{T}_i^{p,q} v - \mathcal{T}_i^{p,\bar{q}} v\|_\infty \le \delta + \gamma H \delta = (1+\gamma H)\delta = H\delta.$$

Subtracting fixed-point equations yields:

$$\|V_i^{p,q} - V_i^{p,\bar{q}}\|_\infty \le \gamma \|V_i^{p,q} - V_i^{p,\bar{q}}\|_\infty + H\delta \implies \|V_i^{p,q} - V_i^{p,\bar{q}}\|_\infty \le H^2\delta.$$

Taking the supremum over unrestricted response policies preserves the perturbation bound for optimal values:

$$\|W_i^q - W_i^{\bar{q}}\|_\infty \le H^2\delta. \tag{64}$$

For any state partition ϕ and ϕ-abstract policy $\sigma \in \Pi_{\phi,i}$, the reverse triangle inequality gives:

$$\left|\|W_i^q - V_i^{\sigma,q}\|_\infty - \|W_i^{\bar{q}} - V_i^{\sigma,\bar{q}}\|_\infty\right| \le \|W_i^q - W_i^{\bar{q}}\|_\infty + \|V_i^{\sigma,q} - V_i^{\sigma,\bar{q}}\|_\infty \le 2H^2\delta.$$

Taking the infimum over $\sigma \in \Pi_{\phi,i}$ establishes that the best-response loss is Lipschitz continuous:

$$|S_q(\phi) - S_{\bar{q}}(\phi)| \le 2H^2\delta. \tag{65}$$

Similarly, the restricted unilateral improvement satisfies $|D_{\phi,q}(p) - D_{\phi,\bar{q}}(p)| \le 2H^2\delta$. Combining these terms gives the unified error-budget inequality:

$$D_i(\pi; q) \le \tau + \varepsilon_{\mathrm{BR}} + (L_S + L_R)\delta + \varepsilon_{\mathrm{plan}} + \varepsilon_{\mathrm{stat}} + \varepsilon_{\mathrm{opt}}, \tag{66}$$

with canonical constants $L_S = L_R = 2H^2$ under the total variation metric.

I Local Certification of Best Responses in Networked Games

When a networked Markov game features a large state space and high induced primal-graph treewidth, computing a global joint abstraction becomes computationally intractable. This appendix formalizes the local certification framework: by constructing a truncated local MDP on a sub-network and bounding boundary interaction leakages, we compute certified upper bounds on the best-response loss without requiring global tree decompositions.

I.1 Factored Markov Games and Truncated Local MDPs

Let the state space factor as $\mathcal{S} = \prod_{j=1}^n \mathcal{S}_j$. Fix a focal player $i \in \mathcal{N}$ facing stationary opponents with joint policy $q \in \Pi_{-i}$. Averaging opponent actions yields the induced single-agent response MDP with stage rewards $r(s,a)$ and transition kernel $P^a(\cdot \mid s)$ for $a \in \mathcal{A}_i$.

For a bounded function $F : \mathcal{S} \to \mathbb{R}$, define the coordinate oscillation:

$$\delta_j(F) := \max_{s,t \in \mathcal{S}: s_{-j} = t_{-j}} |F(s) - F(t)|. \tag{67}$$

We assume the environment satisfies the following spatial sensitivity conditions uniformly over focal actions $a \in \mathcal{A}_i$:

1. **Reward Sensitivity:** There exists a vector $b \in \mathbb{R}_+^n$ such that whenever s, t differ only in coordinate j:
$$|r(s,a) - r(t,a)| \le b_j. \tag{68}$$
2. **Transition Coupling Sensitivity:** There exists a matrix $C \in \mathbb{R}_+^{n \times n}$ such that whenever s, t differ only in coordinate j, there exists a joint coupling (Y, Z) of $P^a(\cdot \mid s)$ and $P^a(\cdot \mid t)$ satisfying:
$$\Pr(Y_k \ne Z_k) \le C_{kj} \quad (\forall k \in [n]). \tag{69}$$

Select a local subset of coordinates $B \subset [n]$ and fix a reference exterior boundary state s^0_{-B}. For local state $x \in \mathcal{S}_B$, let $\iota_B(x) := (x, s^0_{-B}) \in \mathcal{S}$. The truncated local MDP on $\mathcal{S}_B$ is defined by:

$$\bar{r}(x,a) := r(\iota_B(x), a), \qquad \bar{P}^a(y \mid x) := P^a(S'_B = y \mid \iota_B(x)).$$

For a local critic $f : \mathcal{S}_B \to \mathbb{R}$, its canonical lift $F : \mathcal{S} \to \mathbb{R}$ is defined by $F(s) := f(s_B)$. By construction, $\delta_k(F) = \delta_k(f)$ for $k \in B$, and $\delta_k(F) = 0$ for $k \notin B$.

For local critic f and local response policy $\sigma \in \Delta(\mathcal{A}_i)^{\mathcal{S}_B}$, define local Bellman residuals:

$$\alpha_f := \left[\max_{x \in \mathcal{S}_B, a \in \mathcal{A}_i} \left\{\bar{r}(x,a) + \gamma \bar{P}^a f(x) - f(x)\right\}\right]_+, \tag{70}$$

$$\beta_{f,\sigma} := \left[\max_{x \in \mathcal{S}_B} \left\{f(x) - \bar{\mathcal{T}}^\sigma f(x)\right\}\right]_+, \tag{71}$$

and define the boundary influence error:

$$e_B(f) := \sum_{j \notin B} b_j + \gamma \sum_{k \in B} \delta_k(f) \sum_{j \notin B} C_{kj}. \tag{72}$$

Lemma I.1 (Boundary-Residual Transfer)**.** *For every full state $s \in \mathcal{S}$, local state $x = s_B$, and action $a \in \mathcal{A}_i$:*

$$\left|r(s,a) + \gamma P^a F(s) - \left(\bar{r}(x,a) + \gamma \bar{P}^a f(x)\right)\right| \le e_B(f). \tag{73}$$

Consequently, the lifted function F satisfies:

$$\mathcal{T}^* F(s) \le F(s) + \alpha_f + e_B(f), \tag{74}$$

$$\mathcal{T}^\sigma F(s) \ge F(s) - \beta_{f,\sigma} - e_B(f). \tag{75}$$

Proof. Fix state s and let $\bar{s} = \iota_B(s_B)$. Enumerate the exterior coordinates $[n] \setminus B$ as $j_1, \dots, j_m$. Construct a sequence of hybrid states $s^{(0)} = s, s^{(1)}, \dots, s^{(m)} = \bar{s}$ by replacing coordinate j_r with its reference value $s^0_{j_r}$ at step r. Adjacent pairs differ in exactly one coordinate.

Reward discrepancy. Applying reward sensitivity (68) and the triangle inequality gives:

$$|r(s,a) - r(\bar{s},a)| \le \sum_{r=1}^{m} |r(s^{(r-1)},a) - r(s^{(r)},a)| \le \sum_{j \notin B} b_j.$$

Continuation discrepancy. For hybrid states differing only at coordinate j, condition on the coupling (Y, Z) satisfying (69). Because F depends only on s_B:

$$|F(Y) - F(Z)| = |f(Y_B) - f(Z_B)| \le \sum_{k \in B} \delta_k(f) \mathbf{1}\{Y_k \ne Z_k\}.$$

Taking expectations under the coupling yields:

$$|P^a F(s^{(r-1)}) - P^a F(s^{(r)})| \le \sum_{k \in B} \delta_k(f) C_{k,j_r}.$$

Summing over all exterior coordinates $j \notin B$ yields:

$$|P^a F(s) - P^a F(\bar{s})| \le \sum_{k \in B} \delta_k(f) \sum_{j \notin B} C_{kj}.$$

Since $P^a F(\bar{s}) = \bar{P}^a f(s_B)$, scaling by γ and combining with the reward bound establishes (73).

Bellman residuals. By definition of α_f, $\bar{r}(x,a) + \gamma \bar{P}^a f(x) \le f(x) + \alpha_f$. Combining with (73) gives $\mathcal{T}^a F(s) \le F(s) + \alpha_f + e_B(f)$ for all a; maximizing over a proves (74). Averaging over $\sigma(\cdot \mid s_B)$ and applying (71) proves (75). □

I.2 Proof of Proposition 6.1 (Local Best-Response Certificate)

Theorem I.2 (Local Best-Response Certification)**.** *Let W_i^q be the unrestricted optimal response value, and let the local response σ be lifted to the full game by reading s_B. For every state $s \in \mathcal{S}$:*

$$W_i^q(s) \le f(s_B) + H[\alpha_f + e_B(f)], \tag{76}$$

$$V_i^{\sigma,q}(s) \ge f(s_B) - H[\beta_{f,\sigma} + e_B(f)]. \tag{77}$$

Consequently, the lifted policy satisfies:

$$0 \le W_i^q(s) - V_i^{\sigma,q}(s) \le H\left[\alpha_f + \beta_{f,\sigma} + 2e_B(f)\right]. \tag{78}$$

Proof. Define $c_+ := \alpha_f + e_B(f)$. Lemma I.1 establishes $\mathcal{T}^* F \le F + c_+ \mathbf{1}$. Because constant shifts satisfy $\mathcal{T}^*(F + c\mathbf{1}) = \mathcal{T}^* F + \gamma c \mathbf{1}$:

$$\mathcal{T}^*(F + Hc_+\mathbf{1}) = \mathcal{T}^* F + \gamma H c_+ \mathbf{1} \le F + c_+\mathbf{1} + \gamma H c_+ \mathbf{1} = F + Hc_+\mathbf{1},$$

using $1 + \gamma H = H$. Thus, $F + Hc_+\mathbf{1}$ is a Bellman supersolution. Iterating the contraction operator $\mathcal{T}^*$ from this supersolution converges monotonically to the unique fixed point W_i^q, proving $W_i^q \le F + Hc_+\mathbf{1}$, which is (76).

Similarly, define $c_- := \beta_{f,\sigma} + e_B(f)$. By Lemma I.1, $\mathcal{T}^\sigma F \ge F - c_-\mathbf{1}$. Shifting by $-Hc_-\mathbf{1}$ yields:

$$\mathcal{T}^\sigma(F - Hc_-\mathbf{1}) = \mathcal{T}^\sigma F - \gamma H c_- \mathbf{1} \ge F - c_-\mathbf{1} - \gamma H c_- \mathbf{1} = F - Hc_-\mathbf{1}.$$

Thus, $F - Hc_-\mathbf{1}$ is a policy-evaluation subsolution. Iterating $\mathcal{T}^\sigma$ yields a sequence converging monotonically upward to $V_i^{\sigma,q}$, establishing $V_i^{\sigma,q} \ge F - Hc_-\mathbf{1}$, which is (77).

Subtracting (77) from (76) proves (78). □

I.3 Exponential Spatial Decay under Spectral Contraction

Lemma I.3 (Oscillation Recursion)**.** *For every bounded function $F : \mathcal{S} \to \mathbb{R}$, the coordinate oscillation vector $\boldsymbol{\delta}(F) = (\delta_1(F), \ldots, \delta_n(F))$ satisfies:*

$$\boldsymbol{\delta}(\mathcal{T}^* F) \leq b + \gamma \boldsymbol{\delta}(F) C \tag{79}$$

under the row-vector convention.

Proof. Fix states s, t differing only at coordinate j. Let $a_s \in \arg\max_a \mathcal{T}^a F(s)$. Then:

$$\mathcal{T}^* F(s) - \mathcal{T}^* F(t) \leq \mathcal{T}^{a_s} F(s) - \mathcal{T}^{a_s} F(t) \leq b_j + \gamma \sum_{k=1}^{n} \delta_k(F) C_{kj}.$$

Interchanging s and t yields the matching bound for the opposite difference. Maximizing over all pairs (s, t) differing only in coordinate j establishes (79). □

Starting value iteration from $W^{(0)} = 0$ and unrolling (79) yields $\boldsymbol{\delta}(W^{(t)}) \leq b \sum_{\ell=0}^{t-1} \gamma^\ell C^\ell$. When the spectral radius satisfies $\gamma\rho(C) < 1$, the Neumann series converges, establishing:

$$\boldsymbol{\delta}(W_i^q) \leq u := b(I - \gamma C)^{-1}. \tag{80}$$

The local oscillation recursion gives $\boldsymbol{\delta}(W_B) \leq b_B(I - \gamma C_{BB})^{-1} \leq u_B$. Since $u = b + \gamma u C$, summing its exterior coordinates bounds the local boundary contribution. Thus, for a clamped local MDP on subsystem B, the boundary error of the exact local optimal critic W_B satisfies:

$$e_B(W_B) \leq \sum_{j \notin B} u_j. \tag{81}$$

Equip the network coordinates with a shortest-path metric $d(\cdot, \cdot)$ and fix focal agent coordinate i. For spatial decay parameter $a > 0$, define weights $w_j := e^{ad(i,j)}$, maximum amplification factor $\kappa_a := \max_k \sum_j C_{kj} e^{ad(k,j)}$, and aggregate boundary reward $B_a := \sum_j b_j e^{ad(i,j)}$. By the triangle inequality $d(i, j) \leq d(i, k) + d(k, j)$:

$$(Cw)_k = \sum_j C_{kj} e^{ad(i,j)} \leq e^{ad(i,k)} \sum_j C_{kj} e^{ad(k,j)} \leq \kappa_a w_k.$$

Inductively, $C^t w \leq \kappa_a^t w$. Consequently:

$$uw = b \sum_{t=0}^{\infty} \gamma^t C^t w \leq B_a \sum_{t=0}^{\infty} (\gamma\kappa_a)^t = \frac{B_a}{1 - \gamma\kappa_a},$$

provided $\gamma\kappa_a < 1$. For neighborhood ball $B_r(i) := \{j \in [n] \mid d(i, j) \leq r\}$, any exterior node $j \notin B_r(i)$ satisfies $w_j \geq e^{ar}$. Therefore:

$$\sum_{j \notin B_r(i)} u_j \leq e^{-ar} \frac{B_a}{1 - \gamma\kappa_a}.$$

For an exact local solution, $\alpha_f = \beta_{f,\sigma} = 0$. Substituting into Theorem I.2 proves exponential decay of the best-response loss with neighborhood radius r:

$$W_i^q - V_i^{\sigma_r, q} \leq \frac{2HB_a}{1 - \gamma\kappa_a} e^{-ar} \mathbf{1}. \tag{82}$$

I.4 Clipped Operator Bounds under General Nonnegative Coupling

When the spectral contraction condition $\gamma\rho(C) < 1$ fails, the oscillation vector remains bounded by the horizon: $0 \leq \boldsymbol{\delta}(W_i^q) \leq H\mathbf{1}$. Define the clipped operator $\mathcal{F} : [0, H]^n \to [0, H]^n$ componentwise by:

$$\mathcal{F}(u) := \min\{H\mathbf{1},\ b + \gamma u C\}, \qquad u^{(0)} := H\mathbf{1}, \qquad u^{(k+1)} := \mathcal{F}(u^{(k)}). \tag{83}$$

Proposition I.4 (Clipped Monotone Convergence)**.** *The sequence $\{u^{(k)}\}_{k\geq 0}$ is coordinatewise non-increasing, converges to a fixed point $u^\infty = \mathcal{F}(u^\infty)$, and satisfies $\boldsymbol{\delta}(W_i^q) \leq u^{(k)}$ for all $k \geq 0$, whence $\boldsymbol{\delta}(W_i^q) \leq u^\infty$.*

Proof. Because $C \geq 0$ and the minimum operator with a constant vector is non-decreasing, $\mathcal{F}$ is monotone. Since $u^{(1)} = \min\{H\mathbf{1}, b + \gamma H\mathbf{1}C\} \leq H\mathbf{1} = u^{(0)}$, induction establishes that $u^{(k+1)} \leq u^{(k)}$ for all k. Boundedness below by 0 guarantees convergence to fixed point u^∞. From Lemma I.3, $d := \boldsymbol{\delta}(W_i^q) \leq H\mathbf{1}$ and $d \leq b + \gamma dC$, which implies $d \leq \mathcal{F}(d)$. Since $d \leq u^{(0)}$, monotonicity yields $d \leq u^{(k)}$ for all k, completing the proof. □

I.5 Deficit Reformulation and Block-Invariant Subspace Decomposition

Define the deficit variables $q := H\mathbf{1} - b - \gamma H\mathbf{1}C$ and $z^{(k)} := H\mathbf{1} - u^{(k)}$. Substituting into (83) gives:

$$z^{(0)} = 0, \qquad z^{(k+1)} = [q + \gamma z^{(k)} C]_+. \tag{84}$$

Proposition I.5 (Truncation Sensitivity Criterion)**.** *The limit $z^\infty = \lim_{k\to\infty} z^{(k)}$ is the least non-negative fixed point of* (84)*. Moreover:*

$$u^\infty = H\mathbf{1} \iff q \leq 0 \quad \textit{coordinatewise.} \tag{85}$$

Proof. The operator $\mathcal{G}(z) := [q + \gamma zC]_+$ is continuous and monotone. Since $z^{(0)} = 0 \leq z^{(1)}$, the sequence $z^{(k)}$ is non-decreasing in $[0, H]^n$ and converges to the least non-negative fixed point. If $q \leq 0$, then $z^{(1)} = [q]_+ = 0$, whence $z^{(k)} = 0$ for all k, establishing $u^\infty = H\mathbf{1}$. Conversely, if $q_j > 0$ for some coordinate j, then $z_j^{(1)} = q_j > 0$, ensuring $z_j^\infty > 0$ and $u_j^\infty < H$. □

Proposition I.6 (Block-Invariant Subspace Decomposition)**.** *Let the coordinate set partition into an active local block $A \subset [n]$ and an exterior block $A^c = [n] \setminus A$. Suppose:*

$$q_A > 0, \qquad \gamma\rho(C_{AA}) < 1, \qquad q_{A^c} + \gamma q_A (I - \gamma C_{AA})^{-1} C_{A,A^c} \leq 0. \tag{86}$$

Then the fixed point z^∞ has the form:

$$z_A^\infty = q_A (I - \gamma C_{AA})^{-1}, \qquad z_{A^c}^\infty = 0. \tag{87}$$

Proof. Because $C_{AA} \geq 0$ and $\gamma\rho(C_{AA}) < 1$, the resolvent $(I - \gamma C_{AA})^{-1} = \sum_{t=0}^\infty \gamma^t C_{AA}^t$ exists and is non-negative. Set $z_A^* := q_A(I - \gamma C_{AA})^{-1} > 0$ and $z_{A^c}^* := 0$. We prove by induction that $z_{A^c}^{(k)} = 0$ and $0 \leq z_A^{(k)} \leq z_A^*$ for all k.

For the base step $k = 0$, $z^{(0)} = 0$. Assuming the induction hypothesis holds at step k, condition (86) implies:

$$q_{A^c} + \gamma z_A^{(k)} C_{A,A^c} \leq q_{A^c} + \gamma z_A^* C_{A,A^c} \leq 0.$$

Therefore, the positive-part operator sets $z_{A^c}^{(k+1)} = 0$. On the active block A, $q_A > 0$ and all inputs are non-negative, so the positive part is inactive:

$$z_A^{(k+1)} = q_A + \gamma z_A^{(k)} C_{AA} \leq q_A + \gamma z_A^* C_{AA} = z_A^*.$$

Unrolling the recurrence on block A yields $z_A^{(k)} = q_A \sum_{t=0}^{k-1} \gamma^t C_{AA}^t \uparrow q_A(I - \gamma C_{AA})^{-1} = z_A^*$, proving (87). □

Proposition I.6 characterizes the clipped oscillation bound on subsystem A even when the exterior network A^c has high sensitivity.

I.6 Computational Complexity of Local Table Sweeps

Let $L := |\mathcal{S}_B| = \prod_{j \in B} |\mathcal{S}_j|$ denote the local state space size. With local action space size $a := |\mathcal{A}_i|$, the dense transition matrix for the local MDP contains aL^2 entries. Each Bellman evaluation sweep requires $\mathcal{O}(aL^2)$ operations. Value iteration contracts by γ; starting from a value function within sup-norm distance H, reaching target accuracy ε requires $t \geq \frac{\log(H/\varepsilon)}{\log(1/\gamma)}$ iterations. Thus, local response compilation scales as:

$$\mathcal{O}\left(aL^2 \frac{\log(H/\varepsilon)}{\log(1/\gamma)}\right). \tag{88}$$

I.7 Induced-Kernel Sensitivity Matrix Construction

When physical dynamics factorize over coordinates, let E^s_{kj} bound the total variation change of successor coordinate k when current state coordinate j is perturbed under fixed actions. Let $E^a_{k\ell}$ bound the change of successor coordinate k when opponent action ℓ is perturbed under fixed states. Finally, let $\Pi^{(-i)}_{\ell j}$ bound the total variation sensitivity of opponent ℓ's action policy to state coordinate j.

Coupling opponent action selection under maximal coupling and taking a union bound over coordinate changes yields the aggregate sensitivity matrix:

$$C = E^s + E^a \Pi^{(-i)}. \tag{89}$$

Controlling coordinate marginals alone is insufficient for multivariate functions; the joint coupling (Y, Z) in (69) is necessary to ensure valid expectation bounds for non-linear local critics.

I.8 Statistical Estimation Error in Local Certification

When the local Bellman table is estimated from m independent observations per local state-action pair, Hoeffding's inequality over the extreme vertices of the local hypercube $[0, H]^L$ yields empirical confidence radius:

$$\zeta_m := \min\left\{H,\ H\sqrt{\frac{L\log 2 + \log(2L|\mathcal{A}_i|/\alpha)}{2m}}\right\}, \tag{90}$$

holding with probability at least $1 - \alpha$. Combining empirical residuals $\widehat{\alpha}_f, \widehat{\beta}_{f,\sigma}$ with boundary error $e_B(f)$ yields the fully certified finite-sample guarantee:

$$W^q_i - V^{\sigma,q}_i \leq H\left[\widehat{\alpha}_f + \widehat{\beta}_{f,\sigma} + 2e_B(f) + 2\zeta_m\right]\mathbf{1}. \tag{91}$$

J Experimental Protocols, Benchmark Specifications, and Extended Empirical Results

This appendix provides the comprehensive empirical protocols, environment specifications, baseline algorithms, and extended numerical evaluations supporting Section 7. All reported policy losses and Nash gaps are re-evaluated in the original, uncompressed game by an independent evaluator after the abstract state partition ϕ has been fixed.

J.1 Small Finite Discounted Games: Exact Global Optimization

We evaluated Algorithm 1 against exhaustive grid enumeration on a suite of 76 finite discounted games.

Environment Suite. The suite consists of randomly generated two-player and three-player discounted Markov games with state counts $|\mathcal{S}| \in \{3, 4, 5\}$, action spaces $|\mathcal{A}_i| \in \{2, 3\}$, discount factor $\gamma = 0.8$, and fully action-dependent transition matrices. For each game instance, the benchmark opponent pool $\mathcal{Q}$ contains between 2 and 4 stationary opponent profiles.

Verification Protocol. For each instance, we executed two independent solver pipelines:

1. **Exhaustive Grid Search:** We enumerated every possible discrete partition mapping $\phi : \mathcal{S} \to [K]$. For each partition, continuous mixed deployment actions and response heads were searched over the denominator-M grid $\mathcal{G}_M$, computing the exact global discretized optimum c_M by brute force.
2. **Bounded-Treewidth Min–Max Dynamic Programming (Algorithm 1):** We constructed the factor graph $\mathcal{G}_{\mathrm{B}}$, extracted an elimination tree decomposition $\mathcal{T}$, and executed min–max message passing as defined in Equation (7).

Across all 76 instances, the objective value c_M and certificate bounds returned by Algorithm 1 agreed with the brute-force enumeration to machine precision.

Grid Refinement and Numerical Optimization. On a representative three-state action-dependent game with budget $K = 2$, grid resolutions range from $M = 50$ to $M = 1000$. The numerical optimization results give the bracket $[0.050510, 0.050522]$, of width 1.2×10^{-5}. The grid certificate is defined by $L = [Hc_M - H/M]_+$ and the evaluated feasible-system upper endpoint $U = F_{\mathcal{G}}(\widehat{f})$, as in Appendix E.

J.2 Markov Soccer Benchmark under Opponent Distribution Shifts

Environment Specifications. We implement the standard Markov Soccer environment on a 4×5 discrete grid (Ishibashi et al., 2025). The environment consists of two players and a soccer ball. The physical state is defined by the tuple $s = (p_1, p_2, b)$, where $p_1, p_2 \in [20]$ denote the grid positions of Player 1 and Player 2, and $b \in \{1, 2\}$ indicates ball possession. Collisions, ball transfers, and goal boundaries are implemented according to standard rules, yielding 761 total states, 760 playable states, and 1 absorbing terminal goal state. Each player has 5 discrete actions: $\mathcal{A}_i = \{\mathrm{North}, \mathrm{South}, \mathrm{East}, \mathrm{West}, \mathrm{Stand}\}$. The joint action space has size $|\mathcal{A}| = 25$, resulting in 38,000 state-joint-action transition branches. The discount factor is set to $\gamma = 0.9$ ($H = 10$). The full-state reference Nash equilibrium was computed via value iteration, achieving a worst-state Nash gap below 1.6×10^{-9} and Bellman optimality residual below 3×10^{-15}.

Matched-Budget Comparison ($K = 128$). We evaluate five state abstraction algorithms under an identical budget of $K = 128$ abstract states:

1. **On-Policy Predictive Clustering:** Clusters states using k-means on the empirical state visitation and reward features under the reference equilibrium profile.
2. **All-Action Model Similarity:** Aggregates states using bisimulation-style metric clustering based on transition probability distributions $P(\cdot \mid s, \mathbf{a})$ and rewards across all 25 joint actions.
3. **Finite Value Probes:** Clusters states based on evaluated state-action values under 8 fixed policy probe pairs.
4. **Equilibrium-Q Profile:** Clusters states based on centered equilibrium action-value vectors $Q^*(s, \cdot) - V^*(s)\mathbf{1}$.
5. **Best-Response Abstraction (Ours):** Optimizes partition ϕ directly against a benchmark set of $|\mathcal{Q}| = 32$ opponent policies using the joint Bellman formulation.

Table 6: Performance comparison on Markov Soccer under a matched budget of $K = 128$ playable-state labels. Evaluated across $|\mathcal{Q}| = 32$ declared opponent contexts. Lower is better.

Abstraction Method	Best-Response Loss $S_{\mathcal{Q}}(\phi)$	Nash Gap $\mathrm{Gap}_{\mathcal{G}}(\widehat{\pi})$
On-Policy Predictive Clustering	0.447736	0.358626
All-Action Model Similarity	0.618513	0.455954
Finite Value Probes	0.280228	0.287851
Equilibrium-Q Profile	0.420622	0.199557
Best-Response Abstraction (Ours)	**0.108843**	**0.065238**

As reported in Table 6 and Figure 3(a), best-response abstraction achieves a best-response loss of 0.1088, outperforming predictive clustering (0.4477) by 75.7% and value probes (0.2802) by 61.2%. All-action model similarity has loss 0.6185 under the same partition budget.

Growing Opponent Population in League Training. We simulate an expanding league training pool containing 48 diverse opponent strategies (24 per player), generated via checkpointed self-play runs with varying exploration parameters. Starting from an initial pool of 4 opponents, candidate strategies are admitted sequentially. We compare two admission regimes:

- **Adaptive Admission (Ours):** Evaluates the best-response loss $S_q(\phi)$ for all league candidates and admits the opponent profile exhibiting the largest best-response loss. If $S_q(\phi) \leq 0.05$, the partition is reused without modification; otherwise, partition cells with conflicting optimal actions are split.
- **Random Admission Baseline:** Admits opponents in random sequential order, splitting partition cells proportionally.

Across three independent seeds, adaptive opponent selection reached 48 opponents using $K = 240$ abstract states and attained an average held-out best-response loss of 0.0825. In contrast, random insertion paths required an average of 238.3 states but incurred a higher held-out loss of 0.1686 ± 0.0196 (Figure 3(b)). This indicates that best-response abstraction expands representation capacity selectively in response to strategic conflict.

J.3 Scalable Response Compilation on 4-by-4 Connect Four

Environment Scale and Specifications. To evaluate scalability, we benchmarked the scalable response compiler on 4×4 Connect Four. The complete state space contains 161,029 reachable board configurations, comprising 134,289 non-terminal decision states and 304,574 legal directed transitions. Transitions are deterministic. State representations and transition graphs were independently verified using coordinate-array and bitboard bitwise operations.

Benchmark Setup. We generated an opponent pool containing 12 benchmark strategies per player (24 total), trained via minimax depth-limited search with randomized evaluation heuristics across three random seeds. All compared abstraction methods were constrained to preserve the mover identity (Player 1 vs. Player 2 turn) and legal action masks, receiving an identical representation budget of $K = 22{,}057$ abstract states (a $7.3\times$ compression over the 161,029 original states).

Table 7: Matched-budget compilation on 4×4 Connect Four (161,029 states). Storage ratio measures the total memory of the partition index and task response heads normalized by uncompressed deterministic actor tables.

Representation Method	K	Storage Ratio	Best-Response Loss $S_{\mathcal{Q}}(\phi)$	Root Response Loss
Predictive Clustering	22,057	0.3181	0.8222	0.07139
Centered Response-Q Clustering	22,057	0.3181	0.4573	0.06394
Best-Response Refinement (Ours)	**22,057**	**0.3181**	$< 10^{-10}$	**0.00000**

Results and Storage Efficiency. As detailed in Table 7 and Figure 3(c), strategy-complete refinement compiled the full response policy set to numerical precision: best-response loss fell below numerical solver tolerance ($< 10^{-10}$), with zero root response loss. All compared methods retain the same zero-gap deployment. In contrast, predictive clustering and centered response-Q clustering incurred losses of 0.8222 and 0.4573, respectively.

Accounting for the state-to-partition lookup array and 24 task-specific response heads, the compressed strategy-complete representation consumed only 31.81% of the memory required by uncompressed tabular actor tables. The compiled representation preserves the benchmark responses to numerical precision.

J.4 TIC-TAC-TOE PSRO WITH EPISODE-LEVEL OPPONENT MIXTURES

We implemented Tic-Tac-Toe, enumerating all 5,478 reachable board states and 16,167 transitions, with game rules cross-checked against OpenSpiel (Lanctot et al., 2019). To evaluate the population interface developed in Appendix H, opponents were mixed at the episode level. Best responses were computed over public history trees using Bayesian posterior belief states (63).

Under exact policy compilation, the compressed response heads reproduced ground-truth history-conditioned action probabilities exactly, preserving the PSRO empirical meta-game payoff matrix.

Table 8: Exploitability (Root Nash Gap) across PSRO population expansion rounds on 5,478-state Tic-Tac-Toe. Exact response compilation preserves trajectory distributions. The minimax tie-breaking run reaches zero root Nash gap at $t = 2$.

	PSRO Population Expansion Iteration t				
Tie-Breaking Regime	1	2	3	4	5
Minimax Tie-Breaking (Ours)	0.2500	**0.0000**	0.0000	0.0000	0.0000
Lexicographic Tie-Breaking Baseline	0.3333	0.1667	0.1250	0.0417	**0.0000**
Heuristic Action-Sum Baseline	0.5000	0.2500	0.2500	0.1250	0.0833

Table 8 records the convergence trajectory across iterations. With full-game minimax tie-breaking, the strategy-complete PSRO pipeline reached zero root Nash gap after two population expansions ($t = 2$). Conversely, lexicographic tie-breaking required five expansions, while the heuristic action-sum baseline retained a root Nash gap of 0.0833 at the fifth expansion. The trajectories illustrate the sensitivity of the population path to response selection.

J.5 EMPIRICAL MODEL ESTIMATION AND OFFLINE COVERAGE VERIFICATION

To study empirical model estimation in the setting of Appendix G, we conducted empirical model estimation experiments on synthetic discounted Markov games with $|\mathcal{S}| = 8, |\mathcal{A}_i| = 2, \gamma = 0.8$ ($H = 5$).

Fixed Sample Counts. We sampled $m_0 \in \{256, 4096, 65536\}$ transitions per state-joint-action cell, built empirical game $\widehat{\mathcal{G}}$, synthesized partition $\widehat{\phi}$ on $\widehat{\mathcal{G}}$, and evaluated the ground-truth loss $F_{\mathcal{G}}(\widehat{f})$ on $\mathcal{G}$. Table 9 records the empirical model residual, its empirical value-error scale e_{emp}, the ground-truth loss, and the reported empirical coverage rate.

Table 9: Empirical model estimation across sample allocations m_0. Evaluated over 50 independent trials with confidence budget $\alpha = 0.05$. The loss-plus-radius column displays $F_{\mathcal{G}}(\widehat{f}) + e_{\mathrm{emp}}$. The theoretical transfer bound in Appendix G is $F_{\widehat{\mathcal{G}}}(\widehat{f}) + 2e$, with the uniform confidence radius e.

m_0	Empirical Residual β	Radius e_{emp}	Ground Loss	Loss $+e_{\mathrm{emp}}$	Recorded Coverage
256	0.1421	0.7105	0.1834	0.8939	100%
4,096	0.0354	0.1772	0.0812	0.2584	100%
65,536	0.0088	0.0441	0.0521	0.0962	100%

Random Offline Logging Distributions. Under uniform ($\nu_{\min} = 1/D$), moderately skewed ($\nu_{\min} = 0.2/D$), and weak-coverage ($\nu_{\min} = 0.05/D$) behavior policies, sample budgets L satisfying (56) yielded $N_x \geq \frac{L\nu_{\min}}{2}$ in all trials. Model confidence radii use the realized cell counts as in Appendix G; the empirical Bernstein radius β_{safe} in (57) applies to the reward-only specialization with known transitions.

J.6 Local Best-Response Certification on Factored Topologies

To assess the localized best-response certification framework of Appendix I, we implemented factored multi-agent networks featuring homogeneous ring graphs and heterogeneous block architectures with state coordinate counts $n \in \{256, 4096, 65536\}$.

Table 10: Local best-response certification on 1D factored networked games under contraction condition $\gamma\kappa_a < 1$. The local critic is W_B on the radius-r neighborhood. At fixed radius, boundary errors are unchanged across the tested sizes and local solve times remain comparable.

n	Radius r	$\lvert B\rvert$	Boundary Error $e_B(W_B)$	True Best-Response Loss	Solve Time
256	$r = 1$	3	0.1428	0.0812	1.4 ms
256	$r = 2$	5	0.0204	0.0118	4.8 ms
256	$r = 3$	7	0.0029	0.0015	18.2 ms
4,096	$r = 2$	5	0.0204	0.0119	5.1 ms
4,096	$r = 3$	7	0.0029	0.0016	19.0 ms
65,536	$r = 2$	5	0.0204	0.0121	5.4 ms
65,536	$r = 3$	7	0.0029	0.0016	20.3 ms

As summarized in Table 10:

- **Spatial Decay:** On homogeneous networks, boundary error $e_B(W_B)$ decayed by approximately $7\times$ per additional radius unit, consistent with the exponential decay bound $\mathcal{O}(e^{-ar})$ of Equation (82).
- **Scaling with Network Size:** As coordinate count scaled from $n = 256$ to $n = 65{,}536$, local solve time for radius $r = 2$ ranged from 4.8 to 5.4 ms, while the boundary error remained 0.0204.
- **Block-Invariant Verification:** In heterogeneous instances where an exterior supercritical block was attached to a contractive local subsystem, the closed-form stable-core solution (87) matched numerical fixed-point iterations to within 2.1×10^{-11}, consistent with Proposition I.6.

J.7 Implementation Details and Verification

All algorithms were implemented in Python 3.11 using NumPy, SciPy, and CuPy, with mixed-integer linear programming outer relaxations executed using Gurobi 11.0. Exact arithmetic checks were implemented using the Z3 SMT solver over rational fields. Transition graphs and rule engines for soccer and board games were cross-validated across two independent implementations. Source code, random seeds, transition matrices, benchmark opponent checkpoints, and execution scripts are organized in the project repository for full reproducibility.