NASH-PRESERVING SYMMETRY REDUCTIONS FOR MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

The standard solution concept for a Markov game is the Nash equilibrium, a joint policy at which no agent gains by deviating alone. Symmetry can make such a game much smaller, but the smaller game is useful only if an equilibrium found there is still an equilibrium of the original. This is not automatic: a symmetric policy that no agent can improve on within the symmetric class may still be beaten by an asymmetric deviation, so a reduction is trustworthy only if the Nash gap it reports matches the true one. We give two reductions that meet this test exactly. Symmetry among states collapses the game to one on groups of equivalent states; symmetry among interchangeable agents replaces their separate policies with a single shared one. Each preserves the Nash gap against all deviations, not merely symmetric ones, and the two combine when compatible. Because the reduced gap is the true one, we use it in three ways: to tighten existing convergence bounds, to derive a new one for shared policies that exposes the cost of tying agents together, and to cut the samples a generative model needs by pooling symmetric queries. Experiments on cooperative and zero-sum tabular games confirm that the reductions preserve equilibria and accelerate policy optimization.

1 INTRODUCTION

Markov games model sequential interaction among multiple agents (Shapley, 1953; Littman, 1994; Filar & Vrieze, 2012), and their tables of transitions and rewards often repeat strategically identical situations. This redundancy invites a shortcut, restricting attention to policies that respect the symmetry and solving the smaller problem that is reduced from the original one (Ravindran & Barto, 2001; Emmons et al., 2022). In this paper, we target the Nash equilibrium, a standard solution concept in multi-agent learning, at which no agent can gain by changing its own behavior unilaterally. Yet, naively solving the symmetry-reduced problem could break the Nash equilibrium's defining requirement of unilateral deviations: a symmetric policy that no agent can improve on within the symmetric reduction may still be beaten by an asymmetric deviation. Therefore, the question is which symmetric reduction is faithful: one whose best response, taken over all deviations, stays within the reduction itself.

We give two reductions that are faithful in this sense (Section 3), one for each of the two most common symmetries in practice: 1) a group acting on the states so that relabeled states behave identically (Givan et al., 2003; Van der Pol et al., 2020), and 2) the interchangeability of homogeneous agents (Chen et al., 2022a; van der Pol et al., 2022). The two symmetries reduce the game at different levels. A state relabeling leaves every agent in place, so it shrinks the game: the reduction is a smaller Markov game on groups of equivalent states. An agent permutation moves each agent's policy along with the states and actions, so it cannot shrink the game; its reduction is a single policy shared across the interchangeable agents, read through each agent's own view of the state. One mechanism nevertheless proves both faithful. When the other agents play a symmetric policy, the problem that a deviating agent faces inherits the symmetry, so some best response is itself symmetric even though nothing constrains the deviation in advance. We establish this for the state quotient (Section 3.1), then obtain agent sharing from it one agent at a time, applying the quotient result to the permutations that fix the deviating agent's label (Section 3.2); the two reductions combine by the same route when the symmetries commute (Section 3.3). Each preserves the Nash gap against all deviations, not merely symmetric ones.

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.

The distinction between a smaller game and a shared policy resurfaces, and sharpens, when the reductions are turned into algorithmic guarantees (Section 4). The reductions supply no solver, so we specialize to subclasses of Markov games for which Nash equilibrium finding is well understood: Markov potential games, where a single scalar potential tracks every agent's unilateral gains, and two-player zero-sum games, where one agent's gain is the other's loss. Both subclasses have finite-time policy optimization guarantees: independent policy gradient converges globally in the former (Leonardos et al., 2022; Zhang et al., 2024; Chen et al., 2022b) and to Nash in the latter (Daskalakis et al., 2020; Wei et al., 2021). For a state quotient these guarantees transfer at once: the quotient of a potential game is a potential game (Section 4.1) and the quotient of a zero-sum game remains zero-sum (Section 4.3), so existing solvers and generative-model estimators run unchanged on the smaller game, and their bounds carry back to the original with the state count replaced by the orbit count. For agent sharing there is no smaller game to run anything on, and existing analyses do not apply to policies tied across agents. We therefore prove a new convergence rate for policy gradient on the shared table, whose one factor with no quotient analogue is the price of tying agents together (Section 4.1), and design an estimator that pools the queries of interchangeable agents, cutting the number of distinct transition rows to estimate from exponential to polynomial in the number of agents (Section 4.2). Numerical results on game instances (Section 5) validate these claims: the reductions preserve the Nash gap exactly, accelerate policy optimization and reduce sample complexity in proportion to the symmetry exploited.

1.1 Related Work

Prior work on symmetry in sequential decision making preserves values, imposes symmetry as a prior, or proves that symmetric equilibria exist. None has asked whether the Nash gap inside a symmetric reduction equals that of the original game against *all* deviations, symmetric or not.

Single-agent symmetry preserves values. MDP minimization and bisimulation aggregate behaviorally identical states (Givan et al., 2003), symmetry-based minimization makes the acting group explicit (Ravindran & Barto, 2001), and MDP homomorphic networks carry the structure into learned representations (Van der Pol et al., 2020). Our state quotient is the Markov-game form of these constructions, and the value-preservation half of Theorem 1 is theirs. The other half, Eq. (4), has no single-agent counterpart: an optimal policy has no one to deviate from it, whereas a Nash gap is defined by unilateral deviations that need not respect the symmetry.

Symmetry as a prior without equilibrium guarantees. Equivariant actor–critic architectures build permutation and spatial symmetry into policies and critics (van der Pol et al., 2022; Chen & Zhang, 2024), homogeneous agents share an actor with an invariant critic (Chen et al., 2022a), and data augmentation with a soft equivariance penalty exploits approximate symmetry (Yu et al., 2023). These methods use exactly the reductions we study, but their faithfulness is assessed empirically: a policy no agent can improve within the symmetric class may still be beaten by an asymmetric deviation. Our identities supply the missing step, so shared or invariant parameterizations can be trained with the ordinary Nash gap as the target.

Existence results or normal-form reductions. Every finite symmetric game has a symmetric Nash equilibrium (Nash, 1951), but existence of some symmetric equilibrium is weaker than what we need: that the reduced gap of every symmetric policy is its true gap. In symmetric common-payoff games, symmetric local optima are global Nash equilibria (Emmons et al., 2022), and anonymous games reduce a player's view to its own choice and the multiset of the others' (Daskalakis & Papadimitriou, 2007); both concern single-stage (normal-form) games, and our pointed-state orbits are their Markov-game counterpart. Team-symmetric stochastic games share policies within teams and converge empirically to symmetric equilibria (Lee & Hung, 2026), and mean-field methods obtain ϵ-Nash policies through symmetrization and Lipschitz extension in finite-horizon, monotone settings (Yardim & He, 2025); our reductions are exact for discounted, finite-N games without monotonicity, and extend to approximate symmetry. Equilibrium computation in multiplayer and adversarial-team stochastic games (Ganzfried et al., 2019; Kalogiannis et al., 2022) is complementary: we propose no solver, only reductions that shrink the problem any solver faces.

2 PRELIMINARIES

For a finite set $\mathcal{X}$, let $|\mathcal{X}|$ be its cardinality and $\Delta(\mathcal{X})$ its probability simplex. A finite discounted Markov game is specified by $M = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \vec{r}, \mu, \gamma \rangle$, with agents $\mathcal{N} = \{1, \dots, N\}$, finite states $\mathcal{S}$, joint actions $\mathcal{A} = \prod_i \mathcal{A}^i$, transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, rewards vector $\vec{r} = (r^1, \dots, r^N)$ consisting of per-agent rewards $r^i : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and initial state distribution $\mu \in \Delta(\mathcal{S})$, discount $\gamma \in [0, 1)$. We consider product policies $\pi = (\pi^1, \dots, \pi^N)$, with $\pi^i : \mathcal{S} \to \Delta(\mathcal{A}^i)$ being the policy of agent i and $\pi(a|s) = \prod_i \pi^i(a^i|s)$. Let $\pi^{-i} := (\pi^j)_{j \neq i}$ be the policies of all agents other than i. Agent i's value under π is denoted $V^i_\pi(s) := \mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t r^i(s_t, a_t) \,\middle|\, s_0 = s \right]$ and $V^i_\pi(\mu) := \mathbb{E}_{s \sim \mu}[V^i_\pi(s)]$. The Nash gap of a policy π in a Markov game M is

$$\mathrm{NashGap}_M(\pi) := \max_{i \in \mathcal{N}} \max_{\tilde{\pi}^i : \mathcal{S} \to \Delta(\mathcal{A}^i)} \left\{ V^i_{\tilde{\pi}^i, \pi^{-i}}(\mu) - V^i_\pi(\mu) \right\} \tag{1}$$

and π is ϵ-Nash if this quantity is at most ϵ. Unless stated otherwise, every Nash gap below uses this unrestricted stationary deviation class.

3 SYMMETRY REDUCTIONS

Two kinds of symmetry are common in multi-agent problems: a symmetry among states, where relabeling a state changes nothing that matters, and a symmetry among agents, where interchangeable agents can be permuted without changing the game. The two reduce the game differently. A state relabeling leaves each agent in place, so it simply shrinks the game to a smaller Markov game on groups of equivalent states. An agent permutation instead moves each agent's policy along with the states and actions, so it cannot shrink the game itself; its exact reduction is a single decision rule shared across agents, read through each agent's own view of the state.

3.1 FINITE-GROUP STATE QUOTIENTS

A group is a set with an associative product, an identity, and inverses (Dummit & Foote, 2003); it *acts* on a set when each element gives a bijection and the product composes them. A state symmetry is such a group, relabeling states without changing the game's dynamics or rewards. The next definition is not new: it transplants the symmetry conditions of Ravindran & Barto (2001) for MDPs to the Markov-game kernel and reward vector, and the same conditions have been formalized for multi-agent settings, e.g., as $\mathrm{E}(3)$-symmetric cooperative Markov games (Chen & Zhang, 2024) and as multi-agent MDP homomorphisms (van der Pol et al., 2022).

Definition 1. Let a finite group G act on $\mathcal{S}$ by bijections $\{L_g\}_{g \in G}$. M is *G-state-symmetric* if, for all g, s, s', a, i,

$$P(L_g[s'] \mid L_g[s], a) = P(s'|s, a), \qquad r^i(L_g[s], a) = r^i(s, a). \tag{2}$$

Similarly, a policy is G-state-invariant when $\pi^i(a^i \mid L_g[s]) = \pi^i(a^i \mid s)$ for all i, a^i, g, s.

To define the quotient game, write $[s] := \{L_g[s] : g \in G\}$ for the *orbit* of s, $\overline{\mathcal{S}} := \mathcal{S}/G$ for the set of orbits, $\rho(s) := [s]$ for the orbit map, and fix one representative $\iota(\bar{s}) \in \bar{s} \in \overline{\mathcal{S}}$ per orbit.

Definition 2. For a G-state-symmetric M, its *quotient game* $\overline{M} = \langle \mathcal{N}, \overline{\mathcal{S}}, \mathcal{A}, \overline{P}, \vec{\overline{r}}, \overline{\mu}, \gamma \rangle$ has, for any $s \in \bar{s}$,

$$\overline{P}(\bar{s}' \mid \bar{s}, a) := \textstyle\sum_{u \in \bar{s}'} P(u \mid s, a), \qquad \overline{r}^{\,i}(\bar{s}, a) := r^i(s, a), \qquad \overline{\mu}(\bar{s}) := \textstyle\sum_{u \in \bar{s}} \mu(u). \tag{3}$$

For a policy $\bar{\pi}$ in the quotient game, define its lift by $\bar{\pi}^{\uparrow, i}(a^i|s) := \bar{\pi}^i(a^i|\rho(s))$. Let $\Pi(\overline{M})$ be the set of quotient policies. The quotient is well defined: the kernel, rewards, and initial distribution Eq. (3) do not depend on which representative $s \in \bar{s}$ is chosen, since the symmetry conditions Eq. (2) make any two representatives give the same orbit sums. This reduction is the Markov-game form of classical exact state abstraction (Ravindran & Barto, 2001; Givan et al., 2003), and quotient games of this kind already appear in the multi-agent literature (Chen & Zhang, 2024; van der Pol et al., 2022); what is new in Theorem 1 is not the reduction itself but its effect on the equilibrium, which the single-agent setting does not address and the multi-agent work leaves uncertified.

Theorem 1 (Proof in Appendix F)**.** *Let M be a G-state-symmetric Markov game. Then every quotient policy $\bar{\pi} \in \Pi(\overline{M})$ satisfies*

$$\mathrm{NashGap}_M(\bar{\pi}^{\uparrow}) = \mathrm{NashGap}_{\overline{M}}(\bar{\pi}). \tag{4}$$

Proof idea. The identity is proved one agent at a time. Fix a deviating agent i and let the others play the lift $\bar{\pi}^{\uparrow,-i}$, which is G-state-invariant by construction. Averaging the opponents' actions out of P and r^i yields the single-agent MDP that agent i best-responds to; because the opponents' policy is invariant and G relabels only states, leaving action labels fixed, this MDP inherits the symmetry Eq. (2). Its Bellman optimality operator maps G-invariant functions to G-invariant functions, so value iteration from zero converges to an optimal value function that is constant on each orbit, and the same holds for the optimal action values; a greedy policy with a fixed tie-breaking order on $\mathcal{A}^i$ is then an optimal response that is itself G-invariant. The unrestricted maximum in Eq. (1) is therefore attained by an invariant deviation. Invariant local policies on M correspond bijectively to local policies on $\overline{M}$, and lifting preserves values, so agent i's best-response gap in M equals its gap in $\overline{M}$; maximizing over i gives Eq. (4). Unlike the single-agent case, the symmetry is not assumed of the deviating agent: it is inherited from the opponents, which is what lets the quotient certify the gap against deviations that need not respect G.

The content of Eq. (4) is that the reduction is faithful against *unrestricted* deviations, which the fact that lifting preserves values does not give on its own. Optimizing inside the G-invariant class would leave open a profitable *asymmetric* deviation; the identity rules that out, certifying the ordinary Nash gap Eq. (1).

3.2 Agent-Permutation Sharing

State quotienting leaves each agent's policy untouched. Relabeling agents moves states, actions, rewards, and policies together, so the exact reduction is a shared rule rather than a smaller Markov game. Definition 3 belows formalizes the agent permutation symmetry: each agent i carries a local state s^i in a common local state space $\mathcal{C}$ and a local action a^i in a common local action set $\mathcal{B}$, so that $\mathcal{S} = \mathcal{C}^N$ and $\mathcal{A} = \mathcal{B}^N$. Let S_N denote the symmetric group on $\mathcal{N}$ with group action σ acting on states and actions by permuting coordinates, $(\sigma \cdot s)^i := s^{\sigma^{-1}(i)}$ and $(\sigma \cdot a)^i := a^{\sigma^{-1}(i)}$, so that agent $\sigma(i)$ in $\sigma \cdot s$ holds what agent i held in s. The resulting notion of a game with interchangeable agents is standard: it is the homogeneous Markov game of Chen et al. (2022a), and the same permutation conditions underlie permutation-equivariant multi-agent learning (van der Pol et al., 2022; Yu et al., 2023) and team-symmetric stochastic games (Lee & Hung, 2026); its one-shot ancestors are symmetric and anonymous games (Nash, 1951; Daskalakis & Papadimitriou, 2007).

Definition 3. M is *agent-permutation-symmetric* if, for all σ, i, s, s', a,

$$P(\sigma \cdot s' \mid \sigma \cdot s, \sigma \cdot a) = P(s'|s,a), \qquad r^{\sigma(i)}(\sigma \cdot s, \sigma \cdot a) = r^i(s,a). \tag{5}$$

A policy π is *permutation invariant* when $\pi^{\sigma(i)}(b \mid \sigma \cdot s) = \pi^i(b|s)$ for all σ, i, s, b.

Parallel to Section 3.1, the question is whether restricting to $\Pi_{\mathrm{S}_N}(M)$, the set of permutation-invariant product policies, is faithful: i.e., for $\pi \in \Pi_{\mathrm{S}_N}(M)$, whether the unrestricted gap $\mathrm{NashGap}_M(\pi)$ can be computed by maximizing over a reduced class of deviations, as Eq. (4) does over quotient policies. Since there is no quotient game to maximize over, the reduced class must instead be written from the deviating agent's own point of view. A *pointed state* (i, s) is a state with one agent singled out. Permutations act on pointed states by $\sigma \cdot (i, s) := (\sigma(i), \sigma \cdot s)$, and the resulting *pointed orbit* is denoted as $\mathcal{Z} := (\mathcal{N} \times \mathcal{S})/\mathrm{S}_N, z^i(s) := [(i,s)]_{\mathrm{S}_N}$. Concretely, $z^i(s)$ records agent i's own local state together with the multiset of the others', $z^i(s) \;\leftrightarrow\; \big(s^i, \{\{s^k : k \neq i\}\}\big)$, computed by reading off s^i and sorting the other agents' local states.

A mapping $\chi_z : \mathcal{Z} \to \Delta(\mathcal{B})$, read through agent i's own view, $\chi_z^i(b|s) := \chi_z(b|z^i(s))$, is a local policy for agent i. Giving every agent the same table recovers exactly the permutation-invariant policies: $\pi \in \Pi_{\mathrm{S}_N}(M)$ if and only if all agents share the same $\psi : \mathcal{Z} \to \Delta(\mathcal{B})$, i.e.,

$$\pi^i(b|s) = \psi(b|z^i(s)). \tag{6}$$

The reduction thus replaces N policies by one table with $|\mathcal{Z}| = |\mathcal{C}| \binom{|\mathcal{C}|+N-2}{N-1}$ entries instead of the $N\,|\mathcal{C}|^N$ of the full policy table. Theorem 2 below says that a deviating agent's best response may

be taken of the same form χ_z^i, with χ_z now unconstrained. Section 4 builds on the shared-rule form Eq. (6) to prove a convergence rate directly on the single tied table.

Theorem 2 (Proof in Appendices G to I)**.** *Let M be agent-permutation-symmetric and $\pi \in \Pi_{\mathrm{S}_N}(M)$. Then*

$$\mathrm{NashGap}_M(\pi) = \max_{i \in \mathcal{N}} \max_{\chi_z : \mathcal{Z} \to \Delta(\mathcal{B})} \left\{ V^i_{\chi_z^i, \pi^{-i}}(\mu) - V^i_\pi(\mu) \right\}. \tag{7}$$

Proof idea. The identity reduces to state quotienting (Theorem 1), applied one agent at a time with a group tailored to that agent. Fix a deviating agent i and write $H_i := \{\sigma \in \mathrm{S}_N : \sigma(i) = i\}$ for the permutations that keep i's label fixed, its *stabilizer*. Relabeling the other agents by $\tau \in H_i$ moves the state s to $\tau \cdot s$ but leaves agent i's own action untouched; because the opponents' policy π^{-i} is permutation invariant, the transition kernel and reward that agent i faces are unchanged by this relabeling. Agent i's best-response MDP is therefore H_i-state-symmetric, and Theorem 1 (with $G = H_i$ and a single agent) says that some optimal response is constant on each H_i-orbit. Finally, the H_i-orbits are exactly the level sets of z^i: two states are related by a permutation fixing i precisely when agent i holds the same local state in both and the other agents' local states form the same multiset. An H_i-invariant response is thus a function of z^i alone, that is, of the form χ_z^i, which gives Eq. (7) after maximizing over i. The group must depend on i: relabelings that move agent i change which player is deviating, which is why the reduction uses pointed orbits rather than orbits of states.

3.3 Commuting State and Agent Symmetries

When the state and agent actions commute, the two reductions compose into a single reduction to joint orbits, still exact for the unrestricted Nash gap.

Definition 4. M has *commuting symmetries* if it is both G-state-symmetric and agent-permutation-symmetric, and

$$L_g[\sigma \cdot s] = \sigma \cdot L_g[s] \quad \text{for all } g \in G,\ \sigma \in \mathrm{S}_N,\ s \in \mathcal{S}. \tag{8}$$

For such M, $G \times \mathrm{S}_N$ acts on pointed states by $(g, \sigma) \cdot (i, s) = (\sigma(i), L_g[\sigma \cdot s])$. Define the combined pointed orbits

$$\mathcal{W} := (\mathcal{N} \times \mathcal{S})/(G \times \mathrm{S}_N), \qquad w^i(s) := [(i, s)]_{G \times \mathrm{S}_N}, \tag{9}$$

with $\chi_w^i(b|s) := \chi_w(b|w^i(s))$ for $\chi_w : \mathcal{W} \to \Delta(\mathcal{B})$. Call a product policy *jointly invariant* if it is invariant under both actions; these are exactly the policies carried by a single rule $\psi : \mathcal{W} \to \Delta(\mathcal{B})$ through $\pi^i(b|s) = \psi(b|w^i(s))$. A jointly invariant policy is thus one table with $|\mathcal{W}|$ entries, at most the $|\mathcal{Z}|$ of permutation sharing alone. Theorem 3 below says that a deviating agent's best response may be taken of the same form χ_w^i, with χ_w unconstrained, so the Nash gap of such a policy is computed exactly over $\mathcal{W}$. Our proof runs the argument of Theorem 2 with the group $G \times H_i$ in place of the stabilizer H_i.

Theorem 3 (Proof in Appendix M)**.** *Let M have commuting symmetries as in Definition 4. Then every jointly invariant π satisfies*

$$\mathrm{NashGap}_M(\pi) = \max_{i \in \mathcal{N}} \max_{\chi_w : \mathcal{W} \to \Delta(\mathcal{B})} \left\{ V^i_{\chi_w^i, \pi^{-i}}(\mu) - V^i_\pi(\mu) \right\}. \tag{10}$$

4 From Reductions to Guarantees

Each reduction of Section 3 maps its reduced object to a product policy of M whose unrestricted Nash gap equals the reduced gap: the lift $\bar{\pi} \mapsto \bar{\pi}^{\uparrow}$ for state quotients Eq. (4), the shared rule Eq. (6) for agent permutations Eq. (7), and their composition for commuting symmetries Eq. (10). Any solver that drives the reduced gap below ϵ thus yields an ϵ-Nash policy of M. The reductions do not supply a solver, so we specialize to classes where equilibrium finding is well solved: policy optimization in Markov potential games, importing existing bounds and proving a new one for shared parameterizations (Section 4.1); generative-model estimation (Section 4.2); and both for two-player zero-sum games (Section 4.3).

4.1 Policy Optimization in Markov Potential Games

Markov potential games (MPGs) (Monderer & Shapley, 1994; Leonardos et al., 2022), which include fully cooperative games, admit finite-time policy-optimization guarantees because a shared potential tracks every unilateral value change.

Definition 5. M is a *Markov potential game* if a scalar function $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ exists whose discounted total $\Phi_\pi(s) := \mathbb{E}_\pi[\sum_{t\geq 0} \gamma^t \phi(s_t, a_t) \mid s_0 = s]$ tracks every unilateral value change,

$$V^i_{\widetilde{\pi}^i,\pi^{-i}}(s) - V^i_{\pi^i,\pi^{-i}}(s) = \Phi_{\widetilde{\pi}^i,\pi^{-i}}(s) - \Phi_{\pi^i,\pi^{-i}}(s).$$

Cooperative games ($r^i \equiv r$) are MPGs with $\phi = r$.

We fix one representative tabular solver, the *log-barrier softmax policy gradient* of Chen et al. (2022b). Each agent i keeps a table of logits $\theta^i(s,\cdot)$ and plays the softmax $\pi^i_\theta(a^i \mid s) \propto \exp\theta^i(s,a^i)$; all agents simultaneously ascend their own value plus a log-barrier term that repels the iterates from the simplex boundary, and Definition 5 turns these ascents into ascent on the potential. An iterate with Nash gap at most ϵ appears within $T = O\big(N\kappa^2 \left|\mathcal{S}\right|^2 \Delta_\Phi A^2_{\max}/[(1-\gamma)^3\epsilon^2]\big)$ iterations. Here $|\mathcal{S}|$, $A_{\max} := \max_i \left|\mathcal{A}^i\right|$, N, and $(1-\gamma)^{-1}$ measure size; the *potential range* $\Delta_\Phi := \max\Phi_\pi(s) - \min\Phi_\pi(s)$, over product policies and states, scales the objective; and, with $d^\pi_\mu(s) := \sum_{t\geq 0}\gamma^t \mathrm{Pr}^\pi_\mu(s_t = s)$ the discounted visitation from $s_0 \sim \mu$, assumed positive everywhere as in the source, the *distribution mismatch* $\kappa := \max_{\pi,\pi'} \left\| d^\pi_\mu / d^{\pi'}_\mu \right\|_\infty \geq 1$ controls exploration. A symmetry cannot change N, γ, or the action sets, so its gain depends only on how $|\mathcal{S}|$, κ, and Δ_Φ transform.

State quotients. Let M be a G-state-symmetric MPG with potential ϕ. The potential may be taken G-invariant at no cost: the group average $\phi_G(s,a) := \frac{1}{|G|}\sum_{g\in G}\phi(L_g[s],a)$ is again a potential function for M, is G-invariant, and has no larger total-potential range than ϕ. Being constant on orbits, it descends to

$$\overline{\phi}(\bar{s},a) := \phi_G(s,a), \qquad s \in \bar{s}. \tag{11}$$

Proposition 1. *If M is a G-state-symmetric MPG, then $\overline{M}$ is again an MPG with potential $\overline{\phi}$, and positive visitation is inherited from M.*

So the solver runs unchanged on $\overline{M}$, and Eq. (4) makes each iterate's lift an equally good policy of M. Subscripting the constants of $\overline{M}$ by G: $\left|\overline{\mathcal{S}}\right| \leq |\mathcal{S}|$, with $\left|\overline{\mathcal{S}}\right| = |\mathcal{S}| \,/\, |G|$ when G acts *freely* (no nonidentity element fixes a state); $\Delta_{\Phi,G} \leq \Delta_\Phi$, since quotient potentials are potentials of lifted policies; and $\kappa_G \leq \kappa$, since quotient visitation $\overline{d}^{\overline{\pi}}_{\overline{\mu}}(\bar{s}) = \sum_{s\in\bar{s}} d^{\overline{\pi}^\uparrow}_\mu(s)$ sums over orbits and a ratio of sums is bounded by the largest pointwise ratio. The quotient run thus needs $O\big(N\kappa_G^2 \left|\overline{\mathcal{S}}\right|^2 \Delta_{\Phi,G} A^2_{\max}/[(1-\gamma)^3\epsilon^2]\big)$ iterations, at most $|G|^{-2}$ times the count on M under a free action: the state count enters squared, so its exponent becomes the exponent of $|G|$ in the ratio, with the mismatch and potential range contributing further factors at most one. These are ratios of upper bounds, not of running times, and a bound with no explicit state or visitation dependence need not improve; since Eq. (4) constrains only the reduced gap, any other tabular MPG analysis transfers by the same substitution.

Agent permutations. Let M be an agent-permutation-symmetric MPG, so $\mathcal{S} = \mathcal{C}^N$, $\mathcal{A} = \mathcal{B}^N$, and $A_{\max} = |\mathcal{B}|$. There is no reduced game: Eq. (6) ties every agent's policy to one table over the pointed orbits $\mathcal{Z}$, and no existing bound applies to tied coordinates, so we prove a new one. Tie one log-barrier softmax table $\vartheta : \mathcal{Z}\times\mathcal{B} \to \mathbb{R}$ across pointed orbits and run gradient ascent on the tied objective; recall $|\mathcal{Z}| = |\mathcal{C}|\binom{|\mathcal{C}|+N-2}{N-1}$ and write $z_{\max} := \max_{z\in\mathcal{Z}} \left|\{(i,s) : z^i(s) = z\}\right|$ for the *tiling multiplicity*, the most full-table coordinates any single pointed orbit collects.

Corollary 1 (Proof in Appendix L)**.** *Let M be an agent-permutation-symmetric MPG with positive visitation. From $\vartheta_0 = 0$, gradient ascent on the tied log-barrier objective reaches $\min_{0\leq t<T}\mathrm{NashGap}_M(\pi_{\vartheta_t}) \leq \epsilon$ in $T = O\big(z_{\max} N\,\kappa^2 \left|\mathcal{Z}\right|^2 \left|\mathcal{B}\right|^2 \Delta_\Phi/[(1-\gamma)^3\epsilon^2]\big)$ iterations.*

The gap $\mathrm{NashGap}_M$ is the unrestricted one: the proof controls every stationary deviation of every agent, not only deviations of the shared form, through the equivariance of advantages across pointed

orbits that underlies Eq. (7). Against the full-table rate, κ, Δ_Φ, N, and the horizon are unchanged, since sharing restricts only the parameterization, not the game or its visitation. The state count $|\mathcal{S}|^2 = |\mathcal{C}|^{2N}$ becomes $|\mathcal{Z}|^2$, the analogue of $|\mathcal{S}| \to |\overline{\mathcal{S}}|$. What has no analogue is $z_{\max}$, the price of tying: each shared logit receives the summed gradient of every coordinate it ties. Sharing thus improves the bound exactly when $z_{\max}|\mathcal{Z}|^2 < |\mathcal{C}|^{2N}$; as $|\mathcal{Z}| = O(N^{|\mathcal{C}|-1})$ and $z_{\max} \le N|\mathcal{C}|^{N-1}$, this holds for all sufficiently large N but can fail for small N.

4.2 Estimation in Markov Potential Games with a Generative Model

Suppose next the kernel is queried through a generative model returning an independent next-state sample per state–joint-action pair. A certainty-equivalent guarantee estimates every distinct transition row, plans for the potential on the empirical model, and bounds the Nash gap of the resulting policy $\bar{\pi}_{\mathrm{ce}}$, at a cost of (number of distinct rows) $\times$ (per-row budget). Symmetry acts on the first factor: one representative per class of copied rows suffices. Write $\Phi_{\sup} := \max_{\pi,s}|\Phi_\pi(s)|$ and assume $\gamma\Phi_{\sup} > 0$.

State quotients. For G-state-symmetric M, the rows of (s,a) and $(L_g[s],a)$ coincide up to relabeling, so querying one representative $(\iota(\bar{s}),a)$ per orbit n times and recording the empirical frequency of each next-state orbit realizes the generative model of $\overline{M}$ with $|\overline{\mathcal{S}}|\,|\mathcal{A}|$ rows of support $|\overline{\mathcal{S}}|$, and the guarantee on $\overline{M}$ lifts to M by Eq. (4) with $|\mathcal{S}|$ replaced by $|\overline{\mathcal{S}}|$.

Agent permutations. Let M be agent-permutation-symmetric, so $\mathcal{S} = \mathcal{C}^N$ and $\mathcal{A} = \mathcal{B}^N$. There is no quotient game, but the rows of (s,a) and $(\sigma\cdot s, \sigma\cdot a)$ are copies up to relabeling, so the copy classes are the orbits of *pairs*, $\mathcal{Y} := (\mathcal{C}^N \times \mathcal{B}^N)/\mathrm{S}_N$, $y(s,a) := [(s,a)]_{\mathrm{S}_N}$. Choose a representative $\jmath(\bar{y})$ per pair orbit and a permutation $\tau_{s,a}$ taking $\jmath(y(s,a))$ to (s,a); query each representative n times, form its empirical row, and *transport* it to every pair in its orbit by relabeling next states through $\tau_{s,a}$; write $\widetilde{P}$ for the resulting kernel. The row count drops from $|\mathcal{C}|^N|\mathcal{B}|^N$ to $|\mathcal{Y}|$. A pair orbit is a multiset of N local (state, action) pairs, so

$$|\mathcal{Y}| = \binom{|\mathcal{C}||\mathcal{B}|+N-1}{N} \le (N+1)^{|\mathcal{C}||\mathcal{B}|-1} \qquad \text{versus} \qquad |\mathcal{C}|^N|\mathcal{B}|^N, \tag{12}$$

polynomial rather than exponential in the number of agents. Unlike the state quotient, transport leaves the support of each row at $|\mathcal{C}|^N$: the potential MDP still lives on $\mathcal{S}$. Let $\bar{\pi}_{\mathrm{ce}}$ be a certainty-equivalent policy for the potential MDP with kernel $\widetilde{P}$.

Corollary 2 (Proof in Appendix O)**.** *Let M be an agent-permutation-symmetric MPG and fix $\epsilon > 0$, $\delta \in (0,1)$. If $n \ge \frac{8\gamma^2\Phi_{\sup}^2}{(1-\gamma)^2\epsilon^2}\big(|\mathcal{C}|^N \log 2 + \log\frac{|\mathcal{Y}|}{\delta}\big)$, then $\mathrm{NashGap}_M(\bar{\pi}_{\mathrm{ce}}) \le \epsilon$ with probability at least $1-\delta$, using $n|\mathcal{Y}|$ queries in total.*

Estimating every row of M with the same argument needs the same per-row budget for $|\mathcal{C}|^N|\mathcal{B}|^N$ rows, so pooling saves exactly the ratio in Eq. (12). The gain is statistical, not computational: planning still enumerates the $|\mathcal{B}|^N$ joint actions per state. The $|\mathcal{C}|^N \log 2$ term, the generic ℓ_1 per-row price, keeps the *total* budget exponential in N, but it belongs to the estimator, not the reduction: a per-row analysis with logarithmic support dependence (Azar et al., 2012) would combine with the same row set, though whether it survives the dependence transport introduces between copied rows is open. The empirical model need not be an MPG; the argument controls the Nash gap directly.

4.3 Two-Player Zero-Sum Markov Games

Nothing in Section 3 is special to potential games. At the competitive extreme, two-player zero-sum Markov games ($N=2$, $r^1 = -r^2$), the two roles are generally not interchangeable, so the relevant symmetry is a state group. Theorem 1 already transfers the gap of every quotient policy to M; it remains to check that zero-sum solvers apply to $\overline{M}$.

Proposition 2. *If M is a G-state-symmetric two-player zero-sum Markov game, then $\overline{M}$ is also two-player zero-sum.*

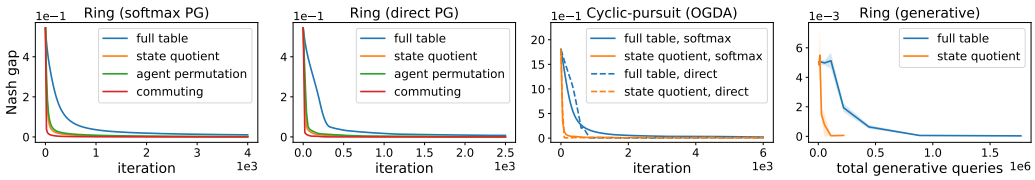


Figure 1: Curves are single deterministic runs from the uniform policy initialization, except the rightmost (generative-model) plot, which shows mean ± standard deviation over 10 seeds.

Optimization. The decentralized *optimistic gradient descent–ascent* (OGDA) of Wei et al. (2021) runs, at every state, a projected optimistic gradient step for each player on a local matrix game built from a slowly updated critic. With exact gradients, its average-gap bound at the largest admissible step size yields an iterate with $\mathrm{NashGap}_M(\pi_t) \le \epsilon$ within $T = \widetilde{O}\big(|\mathcal{S}|^3/[(1-\gamma)^9\epsilon^2]\big)$ iterations. The critic evaluates every state directly, so the bound has no visitation constant or explicit action-count dependence, and the state count is the only factor a symmetry can move. Run on $\overline{M}$ (zero-sum by Proposition 2), OGDA's lifted iterates have the same Nash gap in M by Eq. (4), so the bound holds with $\big|\overline{\mathcal{S}}\big|$ in place of $|\mathcal{S}|$: a factor $|G|^{-3}$ under a free action, since the state count enters cubed.

Estimation. Exactly as in Section 4.2, the pooled estimator realizes the generative model of $\overline{M}$, zero-sum by Proposition 2, with $\big|\overline{\mathcal{S}}\big|\,|\mathcal{A}|$ rows in place of $|\mathcal{S}|\,|\mathcal{A}|$, so the certainty-equivalent guarantee of Zhang et al. (2023), $\widetilde{O}(|\mathcal{S}|\,|\mathcal{A}|\,(1-\gamma)^{-3}\epsilon^{-2})$ queries for an ϵ-Nash policy, holds for M after lifting with $\big|\overline{\mathcal{S}}\big|$ in place of $|\mathcal{S}|$, a factor $|G|^{-1}$ under a free action. That per-row budget depends on support only logarithmically, so the row ratio is the whole saving.

5 Numerical Results

Our experiments aim to answer three questions: (i) do the symmetry reductions accelerate policy optimization without changing the equilibrium, and does the speed-up grow with the amount of symmetry that is shared? (ii) does this benefit carry over to competitive games? (iii) does symmetry lower the sample cost of estimating an unknown model? We use one game from each instantiation; all games are tabular and evaluated by the true (unrestricted) Nash gap, with the optimization runs using exact gradients and the generative-model run estimating from samples.

Environments. *Ring* is the cooperative ring MPG of Section 3.3 with $N = 3$ agents and $m = 8$ sites (commuting symmetry $C_8 \times \mathrm{S}_3$): each agent either stays in place or steps one site clockwise or counterclockwise, succeeding with probability 0.9 and otherwise staying, and the shared reward is the fraction of distinct occupied sites. We also query it through a generative model that returns one next-state sample per state–action pair. *Cyclic-pursuit* is a two-player zero-sum game on a cycle C_{10}: a pursuer and an evader occupy sites (x_p, x_e), each either stays in place or moves deterministically one site clockwise or counterclockwise, and the pursuer earns $r^1 = \mathbf{1}\{x_p = x_e\} = -r^2$ on capture. The two roles are not interchangeable, so its only symmetry is rotation.

Methods. Each reduction is a policy parameterization: it splits the $|\mathcal{S}| \cdot N$ per-state, per-agent policy-table entries into groups that share a single logit vector, and optimizes over those shared logits. The number of groups is the effective dimension, and fewer groups means more sharing. On *Ring* we compare, from least to most symmetric: the *full table*, which shares nothing ($|\mathcal{S}|\cdot N = 1536$ groups); the *state quotient*, which ties states in the same rotation orbit ($\big|\overline{\mathcal{S}}\big| \cdot N = 192$); *agent permutation*, which ties entries with the same pointed orbit in $\mathcal{Z}$ ($|\mathcal{Z}| = 288$); and the *commuting* reduction, which ties by both ($|\mathcal{W}| = 36$); these are the orbit counts of Section 3.3. Tying logits within a group is the same as running the solver on the reduced object, since by value equivalence the tied objective is the reduced one; for the state quotient this is the solver of Section 4.1 run on $\overline{M}$, and for agent permutation it is the tied table of Corollary 1. Each is trained by log-barrier softmax policy gradient, the parameterization of Corollary 1, and by direct (projected) policy gradient. On *Cyclic-pursuit* only the state quotient applies, the relative-position policy $x_e - x_p \bmod 10$ (10 against 100

states), and both players run optimistic gradient descent-ascent (OGDA) (Daskalakis et al., 2020; Wei et al., 2021) under the same two parameterizations. Under the generative model we compare the pooled estimator of Section 4.2, which queries one representative per orbit, against per-state estimation.

Results. On *Ring*, all three symmetric parameterizations converge faster than the full table, in the order commuting $>$ state quotient $>$ agent permutation $>$ full table (Figure 1, first two plots). This tracks the effective dimension: the commuting reduction shares across both the rotation and the agent-permutation symmetry and so keeps the fewest parameters (36 groups), each single symmetry exploits one (192 for the state quotient, 288 for agent permutation), and the full table shares nothing (1536). The ordering agrees with the direction of the rate bounds in Section 4.1, which favor the reduction with the smaller state or pointed-orbit count: here $\left|\overline{\mathcal{S}}\right|^2 = 4096$ against $z_{\max}\left|\mathcal{Z}\right|^2 = 6 \cdot 288^2$, so the bounds also rank the state quotient ahead of agent permutation. Log-barrier softmax and direct policy gradient give the same ordering. On *Cyclic-pursuit*, where only the state quotient applies, it likewise drives the Nash gap to zero several times faster than the full table under both parameterizations (Figure 1, third plot), so the benefit carries to the competitive setting with no zero-sum-specific analysis beyond Proposition 2. Under the generative model the gain reappears in sample cost: across the tested budgets, pooling one representative per orbit reaches each target Nash gap with about $\left|\overline{\mathcal{S}}\right| / |\mathcal{S}| = 1/8$ of the queries per-state estimation needs (Figure 1, fourth plot). The reduction also loses nothing: lifting the state-quotient solution reproduces the full-game Nash gap up to floating-point precision, confirming Theorem 1.

6 Conclusion

A symmetry reduction earns its shortcut only if the Nash gap it reports is the Nash gap of the original game, measured against every unilateral deviation rather than only the symmetric ones. This paper establishes that the two symmetries most common in multi-agent problems pass that test exactly. A state symmetry collapses the game to a quotient game whose Nash gap equals that of the lifted policy; an agent symmetry replaces N policies by one table over pointed orbits whose reduced best-response gap equals the unrestricted one; and when the two symmetries commute they compose. The reductions themselves are classical; the identities are new, and they rest on a single mechanism that has no single-agent analogue: against invariant opponents, a deviating agent's best-response problem inherits the symmetry, so its optimal deviation may be taken symmetric without loss.

Because the reduced gap is the true gap, every guarantee proved on the reduction is a guarantee on the original game. We use this to tighten existing policy-optimization rates for Markov potential games by the size of the quotient, to derive the first convergence rate for a tied log-barrier parameterization that isolates the multiplicity $z_{\max}$ as the exact price of sharing, to cut the queries a generative model needs by pooling symmetric rows, and to carry the same reductions to two-player zero-sum games. Experiments on cooperative and competitive tabular games confirm the picture: lifted quotient solutions reproduce the full-game Nash gap to machine precision, optimization accelerates in proportion to the symmetry exploited, and sample cost falls by the orbit ratio. Symmetry-respecting parameterizations, long used in multi-agent learning as a heuristic prior, can therefore be trained and certified with the ordinary Nash gap as the target.

AI Use Statement

We used generative AI tools to search for and summarize potentially relevant literature, brainstorm directions for the theoretical analysis, and review draft proofs for possible errors or gaps. The authors checked the relevant sources and independently verified the final claims and proofs. The authors take responsibility for all content of this paper.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

Dingyang Chen and Qi Zhang. ${\rm E}(3)$-equivariant actor-critic methods for cooperative multi-agent reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=yShA4VPYZB.

Dingyang Chen, Yile Li, and Qi Zhang. Communication-efficient actor-critic methods for homogeneous markov games. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=xy_2w3J3kH.

Dingyang Chen, Qi Zhang, and Thinh T Doan. Convergence and price of anarchy guarantees of the softmax policy gradient in markov potential games. *arXiv preprint arXiv:2206.07642*, 2022b.

Constantinos Daskalakis and Christos Papadimitriou. Computing equilibria in anonymous games. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 83–93. IEEE, 2007.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33: 5527–5540, 2020.

D.S. Dummit and R.M. Foote. *Abstract Algebra*. Wiley, 2003. ISBN 9780471433347. URL https://books.google.com/books?id=KIGbCgAAQBAJ.

Scott Emmons, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, and Stuart Russell. For learning in symmetric teams, local optima are global nash equilibria. In *International Conference on Machine Learning*, pp. 5924–5943. PMLR, 2022.

J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Business and Economics. Springer New York, 2012. ISBN 9781461240549. URL https://books.google.com/books?id=uXDjBwAAQBAJ.

Sam Ganzfried, Conner Laughlin, and Charles Morefield. Parallel algorithm for approximating nash equilibrium in multiplayer stochastic games with application to naval strategic planning. *arXiv preprint arXiv:1910.00193*, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1):163–223, 2003. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(02)00376-4. URL https://www.sciencedirect.com/science/article/pii/S0004370202003764. Planning with Uncertainty and Incomplete Information.

Fivos Kalogiannis, Ioannis Anagnostides, Ioannis Panageas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Vaggos Chatziafratis, and Stelios Stavroulakis. Efficiently computing nash equilibria in adversarial team markov games. *arXiv preprint arXiv:2208.02204*, 2022.

Duan-Shin Lee and Yu-Hsiu Hung. Delac: A multi-agent reinforcement learning of team-symmetric stochastic games. *arXiv preprint arXiv:2605.12555*, 2026.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gfwON7rAm4.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In William W. Cohen and Haym Hirsh (eds.), *Machine Learning Proceedings 1994*, pp. 157–163. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6. doi: https://doi.org/10.1016/B978-1-55860-335-6.50027-1. URL https://www.sciencedirect.com/science/article/pii/B9781558603356500271.

Dov Monderer and Lloyd S. Shapley. Potential games. 1994. URL https://api.semanticscholar.org/CorpusID:17381677.

John Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, 1951. URL http://jmvidal.cse.sc.edu/library/nash51a.pdf.

Balaraman Ravindran and Andrew G. Barto. Symmetries and model minimization in markov decision processes. 2001. URL https://api.semanticscholar.org/CorpusID:59092891.

Lloyd S. Shapley. Stochastic games*. *Proceedings of the National Academy of Sciences*, 39:1095 – 1100, 1953. URL https://api.semanticscholar.org/CorpusID:263414073.

Elise Van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. Mdp homomorphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4199–4210, 2020.

Elise van der Pol, Herke van Hoof, Frans A Oliehoek, and Max Welling. Multi-agent MDP homomorphic networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=H7HDG--DJF0.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, pp. 4259–4299. PMLR, 2021.

Batuhan Yardim and Niao He. Exploiting approximate symmetry for efficient multi-agent reinforcement learning. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate (eds.), *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 31–44. PMLR, 04–06 Jun 2025. URL https://proceedings.mlr.press/v283/yardim25a.html.

Xin Yu, Rongye Shi, Pu Feng, Yongkai Tian, Jie Luo, and Wenjun Wu. Esp: Exploiting symmetry prior for multi-agent reinforcement learning. In *ECAI 2023: 26th European Conference on Artificial Intelligence, September 30–October 4, 2023, Kraków, Poland–Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, pp. 2946–2953. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2023.

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

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=X1oVDZIABwF.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: Stationary points, convergence, and sample complexity. *IEEE Transactions on Automatic Control*, 69(10):6499–6514, 2024. doi: 10.1109/TAC.2024.3387208.

The appendix follows the order of the main text: it first proves the three reductions of Section 3, then the guarantees of Section 4, and ends with extensions and experimental details.

- **State quotients.** Appendix A shows that the quotient game of Definition 2 is well defined, Appendix B matches invariant policies of M with policies of $\overline{M}$ and shows that their values agree, and Appendix F completes the proof of Theorem 1 by showing that best responses to invariant opponents can be taken invariant.
- **Potential structure.** Appendix C shows that an MPG with a symmetry always has an invariant potential, and Appendix D uses it to prove Proposition 1, including the comparison of the potential range and visitation constants. Appendix E proves the zero-sum counterpart, Proposition 2.
- **Agent permutations.** Appendices G to I prove Theorem 2 in three steps: permutation invariance is equivalent to sharing one rule over pointed orbits, values and visitations are equivariant under agent relabeling, and best responses to shared opponents can themselves be shared.
- **Optimization rates.** Appendix J records the parameterizations and update rules whose existing bounds we transfer, Appendix K gives the game-dependent constants in those bounds and how each changes under the quotient, and Appendix L proves the shared log-barrier rate of Corollary 1.
- **Combined reduction.** Appendix M describes the orbit structure under commuting symmetries and proves Theorem 3.
- **Extensions and estimation.** Appendix N extends Theorem 1 to approximately symmetric games (Theorem N.1), and Appendix O states and proves the generative-model sample-complexity results, Theorem O.1 and Corollaries 2 and O.1.
- **Experiments.** Appendix P lists the hyperparameters behind Section 5.

A Well-Posedness of the Quotient Markov Game

Lemma A.1 (Well-posed quotient)**.** *Let M be a G-state-symmetric Markov game. Then $\overline{P}, \overline{r}, \overline{\mu}$ in Definition 2 are independent of the representative $s \in \bar{s}$, and $\overline{P}(\cdot \mid \bar{s}, a) \in \Delta(\overline{\mathcal{S}})$ for every $(\bar{s}, a)$ with $\overline{\mu} \in \Delta(\overline{\mathcal{S}})$. If in addition M is an MPG and ϕ_G is the invariant potential from Lemma C.1, then $\overline{\phi}$ in Eq.* (11) *is representative independent as well.*

Proof. Let $s, \widetilde{s} \in \bar{s}$. There is a $g \in G$ such that $\widetilde{s} = L_g[s]$. For every next-state orbit $\bar{s}'$, the map $u \mapsto L_g[u]$ is a bijection from $\bar{s}'$ to itself. Reindexing the orbit sum and using transition equivariance,

$$\begin{aligned}\sum_{u \in \bar{s}'} P(u \mid \widetilde{s}, a) &= \sum_{u \in \bar{s}'} P(L_g[u] \mid L_g[s], a) \\ &= \sum_{u \in \bar{s}'} P(u \mid s, a).\end{aligned}$$

Thus Eq. (3) is independent of the representative. Reward and potential invariance similarly give

$$r^i(\widetilde{s}, a) = r^i(s, a), \qquad \phi_G(\widetilde{s}, a) = \phi_G(s, a),$$

so the reward and potential entries of Eq. (3) and Eq. (11) are also representative-independent. The definition of $\overline{\mu}$ already sums over the full orbit and contains no representative choice.

The state orbits partition $\mathcal{S}$. Hence, for any representative $s \in \bar{s}$,

$$\begin{aligned}\sum_{\bar{s}' \in \overline{\mathcal{S}}} \overline{P}(\bar{s}' \mid \bar{s}, a) &= \sum_{\bar{s}' \in \overline{\mathcal{S}}} \sum_{s' \in \bar{s}'} P(s' \mid s, a) \\ &= \sum_{s' \in \mathcal{S}} P(s' \mid s, a) = 1.\end{aligned}$$

Every term is nonnegative, so $\overline{P}(\cdot \mid \bar{s}, a) \in \Delta(\overline{\mathcal{S}})$. Likewise,

$$\sum_{\bar{s} \in \overline{\mathcal{S}}} \overline{\mu}(\bar{s}) = \sum_{\bar{s} \in \overline{\mathcal{S}}} \sum_{s \in \bar{s}} \mu(s) = \sum_{s \in \mathcal{S}} \mu(s) = 1,$$

and $\overline{\mu}$ is nonnegative. Thus $\overline{\mu} \in \Delta(\overline{\mathcal{S}})$. □

B Policy Correspondence and Value Equivalence

For the proofs, define the unnormalized discounted visitation

$$d^{\pi}_{\mu}(s) := \sum_{t\geq 0} \gamma^t \Pr^{\pi}_{\mu}(s_t = s), \tag{13}$$

where $\Pr^{\pi}_{\mu}$ is the trajectory law under π with $s_0 \sim \mu$, and write $\overline{d}^{\bar{\pi}}_{\overline{\mu}}$ for its quotient counterpart. Likewise write $\overline{V}^{i}_{\bar{\pi}}$ and $\overline{Q}^{i}_{\bar{\pi}}$ for the values and action values of $\bar{\pi}$ in $\overline{M}$, and $\Pi_G(M) \subseteq \Pi(M)$ for the G-invariant product policies of M. Define also the action value and advantage

$$Q^{i}_{\pi}(s,a) := \mathbb{E}_{\pi}\left[\sum_{t\geq 0}\gamma^t r^i(s_t,a_t) \,\middle|\, s_0 = s, a_0 = a\right], \qquad A^{i}_{\pi}(s,a) := Q^{i}_{\pi}(s,a) - V^{i}_{\pi}(s).$$

Write $\mathcal{A}^{-i} := \prod_{j\neq i}\mathcal{A}^j$ and $\pi^{-i}(a^{-i} \mid s) := \prod_{j\neq i}\pi^j(a^j \mid s)$ for $a^{-i} \in \mathcal{A}^{-i}$. For a focal action $b \in \mathcal{A}^i$, define the opponents-averaged advantage, written $\mathring{A}$ to distinguish it from the joint-action advantage A,

$$\mathring{A}^{i}_{\pi}(s,b) := \sum_{a^{-i}\in\mathcal{A}^{-i}} \pi^{-i}(a^{-i} \mid s) A^{i}_{\pi}(s,b,a^{-i}). \tag{14}$$

Lemma B.1 (Policy correspondence and value equivalence)**.** *Let M be a G-state-symmetric Markov game and let $\overline{M}$ be its quotient. Project an invariant π by $\pi^{\downarrow,i}(a^i \mid \bar{s}) := \pi^i(a^i \mid \iota(\bar{s}))$. Then the lift $\bar{\pi}^{\uparrow}$ and projection $\pi^{\downarrow}$ satisfy:*

(a) The maps $\bar{\pi} \mapsto \bar{\pi}^{\uparrow}$ and $\pi \mapsto \pi^{\downarrow}$ are mutually inverse bijections between $\Pi(\overline{M})$ and $\Pi_G(M)$.

(b) For every $\bar{\pi} \in \Pi(\overline{M})$, agent i, state s, and joint action a,

$$V^{i}_{\bar{\pi}^{\uparrow}}(s) = \overline{V}^{i}_{\bar{\pi}}(\rho(s)), \qquad\qquad Q^{i}_{\bar{\pi}^{\uparrow}}(s,a) = \overline{Q}^{i}_{\bar{\pi}}(\rho(s),a). \tag{15}$$

Consequently, $V^{i}_{\bar{\pi}^{\uparrow}}(\mu) = \overline{V}^{i}_{\bar{\pi}}(\overline{\mu})$.

(c) For every $\bar{\pi} \in \Pi(\overline{M})$ and $\bar{s} \in \overline{\mathcal{S}}$,

$$\overline{d}^{\bar{\pi}}_{\overline{\mu}}(\bar{s}) = \sum_{s\in\bar{s}} d^{\bar{\pi}^{\uparrow}}_{\mu}(s). \tag{16}$$

Proof. For every $g \in G$, $\rho(L_g[s]) = \rho(s)$. Therefore

$$\bar{\pi}^{\uparrow,i}(a^i \mid L_g[s]) = \bar{\pi}^i(a^i \mid \rho(L_g[s])) = \bar{\pi}^i(a^i \mid \rho(s)) = \bar{\pi}^{\uparrow,i}(a^i \mid s),$$

so every lift is G invariant. The projection is well defined because an invariant policy is constant on each orbit. Direct substitution gives

$$(\bar{\pi}^{\uparrow})^{\downarrow} = \bar{\pi}, \qquad\qquad (\pi^{\downarrow})^{\uparrow} = \pi \qquad \text{for every } \pi \in \Pi_G(M).$$

This proves part (a).

Fix a quotient policy $\bar{\pi}$ and write $\pi = \bar{\pi}^{\uparrow}$. Define

$$\widetilde{V}^i(s) := \overline{V}^{i}_{\bar{\pi}}(\rho(s)).$$

Using $\pi(a \mid s) = \bar{\pi}(a \mid \rho(s))$ and the quotient definitions,

$$\begin{aligned}
\widetilde{V}^i(s) &= \sum_a \bar{\pi}(a \mid \rho(s)) \left[\overline{r}^{\,i}(\rho(s),a) + \gamma \sum_{\bar{s}'} \overline{P}(\bar{s}' \mid \rho(s),a)\overline{V}^{i}_{\bar{\pi}}(\bar{s}')\right] \\
&= \sum_a \pi(a \mid s) \left[r^i(s,a) + \gamma \sum_{\bar{s}'}\sum_{s'\in\bar{s}'} P(s' \mid s,a)\widetilde{V}^i(s')\right] \\
&= \sum_a \pi(a \mid s) \left[r^i(s,a) + \gamma \sum_{s'\in\mathcal{S}} P(s' \mid s,a)\widetilde{V}^i(s')\right].
\end{aligned}$$

Thus $\widetilde{V}^i$ is a fixed point of the Bellman evaluation operator for π on M. That operator is a γ-contraction, so its fixed point is unique and $\widetilde{V}^i = V^i_\pi$. Substituting this value identity into the one-step action-value equations gives

$$\begin{aligned} Q^i_\pi(s,a) &= r^i(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^i_\pi(s') \\ &= \overline{r}^{\,i}(\rho(s),a) + \gamma \sum_{\bar{s}'} \overline{P}(\bar{s}' \mid \rho(s),a) \overline{V}^i_{\bar{\pi}}(\bar{s}') \\ &= \overline{Q}^i_{\bar{\pi}}(\rho(s),a). \end{aligned}$$

This proves Eq. (15). Averaging its value identity over $s \sim \mu$ and grouping by state orbit yields

$$\begin{aligned} V^i_{\bar{\pi}\uparrow}(\mu) &= \sum_{\bar{s}} \sum_{s \in \bar{s}} \mu(s) \overline{V}^i_{\bar{\pi}}(\bar{s}) \\ &= \sum_{\bar{s}} \overline{\mu}(\bar{s}) \overline{V}^i_{\bar{\pi}}(\bar{s}) = \overline{V}^i_{\bar{\pi}}(\overline{\mu}). \end{aligned}$$

For part (c), the state-to-state kernels induced by π and $\bar{\pi}$ satisfy

$$\sum_{s' \in \bar{s}'} P^\pi(s' \mid s) = \overline{P}^{\bar{\pi}}(\bar{s}' \mid \rho(s)). \tag{17}$$

Indeed, this follows by summing $P^\pi(s' \mid s) = \sum_a \pi(a \mid s) P(s' \mid s,a)$ over $\bar{s}'$ and applying Eq. (3). Let $p^\pi_t(s' \mid s_0)$ and $\bar{p}^{\bar{\pi}}_t(\bar{s}' \mid \bar{s}_0)$ denote the corresponding t-step transition probabilities. We claim that

$$\sum_{s' \in \bar{s}'} p^\pi_t(s' \mid s_0) = \bar{p}^{\bar{\pi}}_t(\bar{s}' \mid \rho(s_0)) \qquad \text{for every } t \geq 0. \tag{18}$$

At $t = 0$, both sides are the indicator of $\rho(s_0) = \bar{s}'$. If the identity holds at time t, then grouping the one-step recursion for p^π_{t+1} by the orbit $\bar{u}$ of the intermediate state and using Eq. (17) gives

$$\begin{aligned} \sum_{s' \in \bar{s}'} p^\pi_{t+1}(s' \mid s_0) &= \sum_{\bar{u} \in \overline{\mathcal{S}}} \left(\sum_{u \in \bar{u}} p^\pi_t(u \mid s_0) \right) \overline{P}^{\bar{\pi}}(\bar{s}' \mid \bar{u}) \\ &= \sum_{\bar{u} \in \overline{\mathcal{S}}} \bar{p}^{\bar{\pi}}_t(\bar{u} \mid \rho(s_0)) \overline{P}^{\bar{\pi}}(\bar{s}' \mid \bar{u}) \\ &= \bar{p}^{\bar{\pi}}_{t+1}(\bar{s}' \mid \rho(s_0)). \end{aligned}$$

Thus Eq. (18) follows by induction. Finally,

$$\begin{aligned} \overline{d}^{\bar{\pi}}_{\overline{\mu}}(\bar{s}') &= \sum_{t=0}^{\infty} \gamma^t \sum_{\bar{s}_0} \overline{\mu}(\bar{s}_0) \bar{p}^{\bar{\pi}}_t(\bar{s}' \mid \bar{s}_0) \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{s_0 \in \mathcal{S}} \mu(s_0) \sum_{s' \in \bar{s}'} p^\pi_t(s' \mid s_0) \\ &= \sum_{s' \in \bar{s}'} d^\pi_\mu(s'), \end{aligned}$$

where the second equality uses $\overline{\mu}(\bar{s}_0) = \sum_{s_0 \in \bar{s}_0} \mu(s_0)$ and Eq. (18). This is Eq. (16). □

C Invariant Potentials and Quotient Closure

This appendix states and proves the averaging device used in Section 4.1, namely that potential invariance is a choice of representative rather than an extra assumption.

Lemma C.1 (An invariant potential exists without loss)**.** *Let a finite group H act by symmetries of an MPG on states, on joint actions, and possibly on agent labels, in such a way that local action labels*

are unchanged and only their agent coordinates are permuted. Write $h \cdot i$ for the induced action on agent labels, so $h \cdot i = i$ for a state-only symmetry and $h \cdot i = \sigma(i)$ for an agent permutation. If ϕ is a potential function, then

$$\phi_H(s,a) := \frac{1}{|H|} \sum_{h \in H} \phi(h \cdot s, h \cdot a)$$

is also a potential function and is H invariant. Moreover the total potential $\Phi_\pi^{\phi_H}(s)$ lies, for every product policy π and state s, in the interval spanned by $\{\Phi_{\pi'}^{\phi}(s') : \pi' \in \Pi(M),\, s' \in \mathcal{S}\}$. Consequently ϕ_H has no larger total-potential range than ϕ, no larger $\Phi_{\sup}$, and takes values in $[0,1]$ whenever ϕ does.

The three symmetry classes of Section 3 all satisfy the hypothesis: a state-only group fixes a, an agent permutation acts on action coordinates only by permuting them, and commuting symmetries use $H = G \times \mathrm{S}_N$.

Proof of Lemma C.1. For $h \in H$, set $\phi_h(s,a) := \phi(h \cdot s, h \cdot a)$. Define the relabeled product policy by

$$(h_*\pi)^{h \cdot i}(b \mid h \cdot s) := \pi^i(b|s).$$

This is well defined because the actions on states and agent labels are bijections. Trajectories under π from s, after applying h to states, joint actions, and agent labels, therefore have the law of trajectories under $h_*\pi$ from $h \cdot s$. The symmetry identities give

$$\Phi_\pi^{\phi_h}(s) = \Phi_{h_*\pi}^{\phi}(h \cdot s), \qquad V_\pi^i(s) = V_{h_*\pi}^{h \cdot i}(h \cdot s). \tag{19}$$

If two profiles differ only in agent i's policy, their relabelings differ only in agent $h \cdot i$'s policy. Applying the MPG identity for ϕ to the relabeled profiles in Eq. (19), and then using reward symmetry, shows that the change in Φ^{ϕ_h} equals agent i's value change. Thus every ϕ_h is a potential function. The MPG identity is linear in the potential function, so their average ϕ_H is also a potential.

For $k \in H$, reindexing h by hk gives

$$\phi_H(k \cdot s, k \cdot a) = \frac{1}{|H|} \sum_{h \in H} \phi(hk \cdot s, hk \cdot a) = \phi_H(s,a).$$

Finally, by linearity of Φ in the potential function and the first identity in Eq. (19),

$$\Phi_\pi^{\phi_H}(s) = \frac{1}{|H|} \sum_{h \in H} \Phi_{h_*\pi}^{\phi}(h \cdot s),$$

an average of values of Φ^ϕ at product policies, since each $h_*\pi$ is again a product policy. Hence $\Phi_\pi^{\phi_H}(s)$ lies in the interval spanned by Φ^ϕ over product policies and states. Taking maxima and minima over π and s gives the range and $\Phi_{\sup}$ claims, and the range statement applied pointwise to the potential function itself gives $\phi_H \in [0,1]$ when $\phi \in [0,1]$, because ϕ_H is an average of values of ϕ. □

D THE QUOTIENT IS A MARKOV POTENTIAL GAME

Lemma D.1 (Inherited MPG structure and quantitative constants)**.** *Let M be a G-state-symmetric MPG and choose the invariant potential ϕ_G from Lemma C.1. Then:*

(a) $\overline{M}$ is an MPG with potential $\overline{\phi}$, and $\Delta_{\Phi,\overline{M}} \le \Delta_{\Phi,M}$ and $\Phi_{\sup,\overline{M}} \le \Phi_{\sup,M}$.

(b) If every product policy of M has positive discounted visitation at every state, then the same is true in $\overline{M}$.

(c) Under the condition in part (b), the quotient mismatch coefficient satisfies $\kappa_G \le \kappa$.

Proof. Throughout this proof, let Φ denote the total potential generated by ϕ_G. Let $\overline{\Phi}_{\bar{\pi}}$ be the discounted total potential generated by $\overline{\phi}$ in $\overline{M}$. The Bellman argument in Lemma B.1(b), with ϕ_G and $\overline{\phi}$ in place of r^i and $\overline{r}^{\,i}$, gives

$$\Phi_{\bar{\pi}^\uparrow}(s) = \overline{\Phi}_{\bar{\pi}}(\rho(s)). \tag{20}$$

This total-potential identity underlies Proposition 1. Fix an agent i, a quotient policy $\bar{\pi}$, and an alternative local policy $\bar{\nu}^i$. Set $\widetilde{\bar{\pi}} := (\bar{\nu}^i, \bar{\pi}^{-i})$. Lifting acts coordinatewise, so $\widetilde{\bar{\pi}}^{\,\uparrow}$ and $\bar{\pi}^\uparrow$ differ only in agent i's policy. At any representative $s \in \bar{s}$, value equivalence, the MPG identity of M, and Eq. (20) give

$$\begin{aligned}
\overline{V}^i_{\widetilde{\bar{\pi}}}(\bar{s}) - \overline{V}^i_{\bar{\pi}}(\bar{s}) &= V^i_{\widetilde{\bar{\pi}}^{\,\uparrow}}(s) - V^i_{\bar{\pi}^\uparrow}(s) \\
&= \Phi_{\widetilde{\bar{\pi}}^{\,\uparrow}}(s) - \Phi_{\bar{\pi}^\uparrow}(s) \\
&= \overline{\Phi}_{\widetilde{\bar{\pi}}}(\bar{s}) - \overline{\Phi}_{\bar{\pi}}(\bar{s}).
\end{aligned}$$

Hence $\overline{\phi}$ is a potential for $\overline{M}$. Moreover, Eq. (20) identifies quotient potential values with the subset generated by lifted policies. Therefore the quotient maximum cannot increase and its minimum cannot decrease, proving the two range inequalities in part (a).

By Eq. (16),

$$\overline{d}^{\bar{\pi}}_{\overline{\mu}}(\bar{s}) = \sum_{s \in \bar{s}} d^{\bar{\pi}^\uparrow}_{\mu}(s).$$

If full-game visitation is positive, every sum is positive, proving part (b). For quotient policies $\bar{\pi}, \bar{\pi}'$ and a quotient state $\bar{s}$, the mediant inequality $\big(\sum_s a_s\big)/\big(\sum_s b_s\big) \le \max_s a_s/b_s$, valid for positive b_s, gives

$$\begin{aligned}
\frac{\overline{d}^{\bar{\pi}}_{\overline{\mu}}(\bar{s})}{\overline{d}^{\bar{\pi}'}_{\overline{\mu}}(\bar{s})} &= \frac{\sum_{s \in \bar{s}} d^{\bar{\pi}^\uparrow}_{\mu}(s)}{\sum_{s \in \bar{s}} d^{\bar{\pi}'^\uparrow}_{\mu}(s)} \\
&\le \max_{s \in \bar{s}} \frac{d^{\bar{\pi}^\uparrow}_{\mu}(s)}{d^{\bar{\pi}'^\uparrow}_{\mu}(s)} \le \kappa.
\end{aligned}$$

Maximizing proves part (c). □

E Proof of Proposition 2

Proof of Proposition 2. Let M be G-state-symmetric and two-player zero-sum, so $N = 2$ and $r^1(s,a) = -r^2(s,a)$ for all (s,a). By Definition 2, the quotient has the same two players and the same joint action set, and its rewards are $\overline{r}^{\,i}(\bar{s},a) = r^i(s,a)$ for any representative $s \in \bar{s}$, which is well defined because reward invariance Eq. (2) makes r^i constant on each orbit. Hence for every orbit $\bar{s}$ and joint action a, with $s \in \bar{s}$ arbitrary,

$$\overline{r}^{\,1}(\bar{s},a) = r^1(s,a) = -r^2(s,a) = -\overline{r}^{\,2}(\bar{s},a),$$

so $\overline{M}$ is two-player zero-sum. Nothing else in the definition of the class involves the kernel or the initial distribution, so no further condition needs checking. □

F Nash-Gap Preservation

The key point is that, against invariant opponents, an agent has an invariant optimal best response even though the original Nash gap allows all local policies.

Lemma F.1 (Invariant optimal best response)**.** *Let M be a G-state-symmetric Markov game. Fix an agent i and a G-invariant product policy π. The best-response MDP faced by agent i when the other agents play π^{-i} is G-state-symmetric and has a G-invariant optimal stationary policy. Consequently,*

$$\max_{\widetilde{\pi}^i} V^i_{\widetilde{\pi}^i, \pi^{-i}}(\mu) = \max_{\widetilde{\pi}^i \ G\text{-invariant}} V^i_{\widetilde{\pi}^i, \pi^{-i}}(\mu). \tag{21}$$

Proof. The best-response MDP has state space $\mathcal{S}$, action space $\mathcal{A}^i$, discount γ, initial distribution μ, and transition kernel and reward

$$P^i_{\pi^{-i}}(s' \mid s, a^i) := \sum_{a^{-i}} \pi^{-i}(a^{-i} \mid s) P(s' \mid s, a^i, a^{-i}),$$
$$r^i_{\pi^{-i}}(s, a^i) := \sum_{a^{-i}} \pi^{-i}(a^{-i} \mid s) r^i(s, a^i, a^{-i}).$$

Since π^{-i} is invariant and the group leaves every action label fixed,

$$\begin{aligned} P^i_{\pi^{-i}}(L_g[s'] \mid L_g[s], a^i) &= \sum_{a^{-i}} \pi^{-i}(a^{-i} \mid L_g[s]) P(L_g[s'] \mid L_g[s], a^i, a^{-i}) \\ &= \sum_{a^{-i}} \pi^{-i}(a^{-i} \mid s) P(s' \mid s, a^i, a^{-i}) \\ &= P^i_{\pi^{-i}}(s' \mid s, a^i), \end{aligned}$$

where the second equality uses policy invariance and Eq. (2). The same calculation with rewards gives

$$r^i_{\pi^{-i}}(L_g[s], a^i) = r^i_{\pi^{-i}}(s, a^i).$$

Thus the best-response MDP is G-state-symmetric.

Let $\mathcal{T}$ be its Bellman optimality operator. If a function V is G invariant, then transition and reward symmetry imply

$$(\mathcal{T}V)(L_g[s]) = (\mathcal{T}V)(s).$$

Starting value iteration from the zero function therefore produces only invariant functions. Because $\mathcal{T}$ is a γ-contraction, these iterates converge to the unique optimal value V^*, which is invariant. Its optimal action-value function consequently satisfies

$$Q^*(L_g[s], a^i) = Q^*(s, a^i).$$

Choose a greedy deterministic policy and break ties using one fixed total order on $\mathcal{A}^i$ at every state. The action-value vectors are identical within each orbit, so the selected greedy action is also identical within each orbit. This gives a G-invariant optimal stationary policy. Since this policy attains the unrestricted best-response value, Eq. (21) follows. □

Lemma F.2 (Nash-gap preservation)**.** *Let M be a G-state-symmetric Markov game. Every quotient policy $\bar{\pi}$ satisfies*

$$\mathrm{NashGap}_M(\bar{\pi}^{\uparrow}) = \mathrm{NashGap}_{\overline{M}}(\bar{\pi}).$$

Consequently, $\bar{\pi}$ is ϵ-Nash in $\overline{M}$ if and only if $\bar{\pi}^{\uparrow}$ is ϵ-Nash in M.

Proof. Write $\pi = \bar{\pi}^{\uparrow}$. Fix an agent i. Since π^{-i} is invariant, Lemma F.1 allows the original best-response optimization to be restricted to invariant local deviations:

$$\max_{\widetilde{\pi}^i} V^i_{\widetilde{\pi}^i, \pi^{-i}}(\mu) = \max_{\widetilde{\pi}^i \ G\text{-invariant}} V^i_{\widetilde{\pi}^i, \pi^{-i}}(\mu).$$

By Lemma B.1(a), projection is a bijection between invariant local policies on M and local policies on $\overline{M}$. Concretely, for an invariant deviation $\widetilde{\pi}^i$, define

$$\bar{\nu}^i(a^i \mid \bar{s}) := \widetilde{\pi}^i(a^i \mid \iota(\bar{s})).$$

Then $(\widetilde{\pi}^i, \pi^{-i}) = (\bar{\nu}^i, \bar{\pi}^{-i})^{\uparrow}$. Value equivalence gives

$$V^i_{\widetilde{\pi}^i, \pi^{-i}}(\mu) = \overline{V}^i_{\bar{\nu}^i, \bar{\pi}^{-i}}(\overline{\mu}),$$
$$V^i_{\pi}(\mu) = \overline{V}^i_{\bar{\pi}}(\overline{\mu}).$$

Therefore,

$$\max_{\widetilde{\pi}^i} V^i_{\widetilde{\pi}^i, \pi^{-i}}(\mu) - V^i_{\pi}(\mu) = \max_{\bar{\nu}^i} \overline{V}^i_{\bar{\nu}^i, \bar{\pi}^{-i}}(\overline{\mu}) - \overline{V}^i_{\bar{\pi}}(\overline{\mu}).$$

The left and right sides are respectively agent i's contributions to the Nash gaps in M and $\overline{M}$. Maximizing over i proves the identity in Eq. (4). □

G Permutation Invariance and Shared Pointed-State Policies

Lemma G.1 (Pointed-state orbits and shared policies)**.** *For any S_N-action on $\mathcal{S}$, a product policy is permutation invariant if and only if it has the shared form Eq.* (6)*. In the canonical model $\mathcal{S} = \mathcal{C}^N$, the equality $z^i(s) = z^j(s')$ is equivalent to*

$$s^i = s'^j, \qquad \{\{s^k : k \neq i\}\} = \{\{s'^k : k \neq j\}\}.$$

Proof. By definition,

$$z^i(s) = z^j(s') \quad \Longleftrightarrow \quad \text{some } \sigma \in \mathrm{S}_N \text{ satisfies } j = \sigma(i),\ s' = \sigma \cdot s. \tag{22}$$

Suppose first that $\mathcal{S} = \mathcal{C}^N$ with the canonical action. The forward direction of Eq. (22) gives

$$s'^j = (\sigma \cdot s)^{\sigma(i)} = s^i,$$

and the coordinates of s' other than j are a reordering of those of s other than i.

Conversely, suppose the two local-state-and-multiset pairs agree. Set $\sigma(i) = j$. Equality of their multisets gives a bijection

$$\sigma : \mathcal{N} \setminus \{i\} \longrightarrow \mathcal{N} \setminus \{j\}$$

such that $s'^{\sigma(k)} = s^k$ for every $k \neq i$. Together with $s'^j = s^i$, this defines a permutation of $\mathcal{N}$ satisfying

$$(\sigma \cdot s)^{\sigma(k)} = s^k = s'^{\sigma(k)} \qquad \text{for every } k \in \mathcal{N}.$$

Thus $\sigma \cdot s = s'$ and $\sigma \cdot (i, s) = (j, s')$, proving the concrete characterization.

Now let π be permutation invariant. If $z^i(s) = z^j(s')$, choose the witness σ from Eq. (22). Then permutation invariance of π gives

$$\pi^j(b \mid s') = \pi^{\sigma(i)}(b \mid \sigma \cdot s) = \pi^i(b|s).$$

Consequently,

$$\psi(b \mid z) := \pi^i(b|s) \qquad \text{for any } (i, s) \text{ such that } z^i(s) = z$$

is well defined and gives Eq. (6).

Conversely, suppose Eq. (6) holds. For every σ,

$$z^{\sigma(i)}(\sigma \cdot s) = z^i(s),$$

because the two pointed states are in the same orbit. Therefore

$$\pi^{\sigma(i)}(b \mid \sigma \cdot s) = \psi\Big(b \mid z^{\sigma(i)}(\sigma \cdot s)\Big) = \psi(b|z^i(s)) = \pi^i(b|s),$$

so π is permutation invariant.

Every $z \in \mathcal{Z}$ is the orbit of some pointed state. It follows that the shared rule inducing an invariant policy is unique, so the correspondence is a bijection. Finally, per-agent invariance implies the joint identity

$$\pi(\sigma \cdot a \mid \sigma \cdot s) = \pi(a \mid s), \tag{23}$$

because

$$\pi(\sigma \cdot a \mid \sigma \cdot s) = \prod_{i \in \mathcal{N}} \pi^{\sigma(i)}(a^i \mid \sigma \cdot s) = \prod_{i \in \mathcal{N}} \pi^i(a^i \mid s).$$

□

H VALUE, ADVANTAGE, VISITATION, AND POTENTIAL EQUIVARIANCE

Lemma H.1 (Strategic equivariance under agent relabeling)**.** *Let M be agent-permutation-symmetric. Every $\pi \in \Pi_{\mathrm{S}_N}(M)$ satisfies, for every $\sigma \in \mathrm{S}_N$, $i \in \mathcal{N}$, $s \in \mathcal{S}$, $a \in \mathcal{B}^N$, and $b \in \mathcal{B}$,*

$$V_\pi^{\sigma(i)}(\sigma \cdot s) = V_\pi^i(s), \qquad Q_\pi^{\sigma(i)}(\sigma \cdot s, \sigma \cdot a) = Q_\pi^i(s,a), \tag{24}$$
$$\mathring{A}_\pi^{\sigma(i)}(\sigma \cdot s, b) = \mathring{A}_\pi^i(s,b).$$

Call μ exchangeable *if $\mu(\sigma \cdot s) = \mu(s)$ for all σ, s. If μ is exchangeable, then $d_\mu^\pi(\sigma \cdot s) = d_\mu^\pi(s)$. If the potential function is permutation invariant, then $\Phi_\pi(\sigma \cdot s) = \Phi_\pi(s)$. In particular, there is a well-defined function $\widehat{\mathring{A}}_\pi$ such that*

$$\mathring{A}_\pi^i(s,b) = \widehat{\mathring{A}}_\pi(z^i(s), b). \tag{25}$$

Proof. First consider the state kernel induced by π:

$$P^\pi(s' \mid s) := \sum_{a \in \mathcal{B}^N} \pi(a \mid s) P(s' \mid s, a).$$

Reindexing the action sum by $a' = \sigma \cdot a$ and using Eq. (23) and Eq. (5) gives

$$\begin{aligned} P^\pi(\sigma \cdot s' \mid \sigma \cdot s) &= \sum_{a \in \mathcal{B}^N} \pi(\sigma \cdot a \mid \sigma \cdot s) P(\sigma \cdot s' \mid \sigma \cdot s, \sigma \cdot a) \\ &= P^\pi(s' \mid s). \end{aligned} \tag{26}$$

The map

$$(s_t, a_t)_{t \geq 0} \longmapsto (\sigma \cdot s_t, \sigma \cdot a_t)_{t \geq 0}$$

is therefore a probability-preserving bijection between trajectories under π started from s and trajectories under π started from $\sigma \cdot s$. By Eq. (5), the reward of agent $\sigma(i)$ along the transformed trajectory equals the reward of agent i along the original trajectory at every time. It follows that

$$V_\pi^{\sigma(i)}(\sigma \cdot s) = V_\pi^i(s).$$

Applying the same reindexing to the one-step action-value equations gives

$$Q_\pi^{\sigma(i)}(\sigma \cdot s, \sigma \cdot a) = Q_\pi^i(s,a).$$

If ϕ is permutation invariant, the same trajectory argument with ϕ in place of r^i proves

$$\Phi_\pi(\sigma \cdot s) = \Phi_\pi(s).$$

For the visitation claim, suppose μ is exchangeable and let $p_t^\pi(s' \mid s)$ be the t-step transition probability under P^π. Equation (26) implies by induction that

$$p_t^\pi(\sigma \cdot s' \mid \sigma \cdot s) = p_t^\pi(s' \mid s) \qquad \text{for every } t \geq 0.$$

Using the exchangeability of μ and reindexing the initial state as $s_0 = \sigma \cdot u$,

$$\begin{aligned} d_\mu^\pi(\sigma \cdot s') &= \sum_{t=0}^\infty \gamma^t \sum_{s_0 \in \mathcal{S}} \mu(s_0) p_t^\pi(\sigma \cdot s' \mid s_0) \\ &= \sum_{t=0}^\infty \gamma^t \sum_{u \in \mathcal{S}} \mu(\sigma \cdot u) p_t^\pi(\sigma \cdot s' \mid \sigma \cdot u) \\ &= d_\mu^\pi(s'). \end{aligned}$$

Subtracting the value identity from the action-value identity gives joint advantage equivariance:

$$A_\pi^{\sigma(i)}(\sigma \cdot s, \sigma \cdot a) = A_\pi^i(s,a). \tag{27}$$

Fix $b \in \mathcal{B}$ and reindex the other agents' actions by $a' = \sigma \cdot a$ with $a^i = b$. Since $(\sigma \cdot a)^{\sigma(i)} = b$, product-policy invariance gives

$$\pi^{-\sigma(i)}\Big((\sigma \cdot a)^{-\sigma(i)} \,\Big|\, \sigma \cdot s\Big) = \pi^{-i}(a^{-i} \mid s).$$

Applying this identity and Eq. (27) in the definition Eq. (14) yields

$$\mathring{A}^{\sigma(i)}_{\pi}(\sigma \cdot s, b) = \mathring{A}^{i}_{\pi}(s, b).$$

Finally, if $z^i(s) = z^j(s')$, then Lemma G.1 supplies a σ with $j = \sigma(i)$ and $s' = \sigma \cdot s$. The local-advantage identity therefore gives

$$\mathring{A}^{j}_{\pi}(s', b) = \mathring{A}^{i}_{\pi}(s, b).$$

Thus defining $\widehat{\mathring{A}}_{\pi}(z, b)$ using any pointed state in orbit z is independent of the choice and proves Eq. (25). □

I Pointed-Orbit Best Responses and the Unrestricted Nash Gap

Lemma I.1 (Pointed-orbit optimal best response)**.** *Let M be agent-permutation-symmetric and $\pi \in \Pi_{\mathrm{S}_N}(M)$. For every agent i, the unrestricted best response reduces to shared rules,*

$$\max_{\nu^i : \mathcal{S} \to \Delta(\mathcal{B})} V^{i}_{\nu^i, \pi^{-i}}(\mu) = \max_{\chi_z : \mathcal{Z} \to \Delta(\mathcal{B})} V^{i}_{\chi^i_z, \pi^{-i}}(\mu). \tag{28}$$

Consequently, Eq. (7) *holds. If μ is exchangeable, all agents have the same best-response gain.*

Proof. Fix an agent i and recall its stabilizer $H_i = \{\tau \in \mathrm{S}_N : \tau(i) = i\}$ from Section 3.2. Against π^{-i}, agent i faces the finite discounted MDP with state space $\mathcal{S}$, action space $\mathcal{B}$, and

$$P^{i}_{\pi^{-i}}(s' \mid s, b) := \sum_{a^{-i}} \pi^{-i}(a^{-i} \mid s) P(s' \mid s, b, a^{-i}),$$
$$r^{i}_{\pi^{-i}}(s, b) := \sum_{a^{-i}} \pi^{-i}(a^{-i} \mid s) r^{i}(s, b, a^{-i}).$$

For $\tau \in H_i$, reindex the sum over a^{-i} by the permutation of the other agents' action coordinates induced by τ. Because $\tau(i) = i$, the focal action b is unchanged. Permutation invariance of π gives

$$\pi^{-i}\big((\tau \cdot a)^{-i} \,\big|\, \tau \cdot s\big) = \pi^{-i}(a^{-i} \mid s) \qquad \text{when } a^i = b.$$

Combining this identity with the two parts of Eq. (5), and using $\tau(i) = i$, therefore gives

$$P^{i}_{\pi^{-i}}(\tau \cdot s' \mid \tau \cdot s, b) = P^{i}_{\pi^{-i}}(s' \mid s, b), \tag{29}$$
$$r^{i}_{\pi^{-i}}(\tau \cdot s, b) = r^{i}_{\pi^{-i}}(s, b). \tag{30}$$

Thus the best-response MDP is H_i-state-symmetric with unchanged action labels.

Its Bellman optimality operator maps H_i-invariant value functions to H_i-invariant value functions by Equations (29) and (30). Starting value iteration from zero and using contraction therefore shows that the optimal value and optimal action-value functions are H_i invariant. Choosing a greedy action with one fixed tie-breaking order on $\mathcal{B}$ gives an H_i-invariant optimal stationary policy.

The H_i-orbits in $\mathcal{S}$ are exactly the fibers of z^i: by definition, $z^i(s) = z^i(s')$ if and only if some permutation fixing i maps s to s'. Hence an H_i-invariant local policy is exactly a policy of the form χ^i_z. The invariant optimal policy constructed above proves Eq. (28). Maximizing that identity over i and subtracting each baseline value proves Eq. (7).

It remains only to prove that, under exchangeability, all agents have the same best-response gain. Fix i, j and choose $\sigma \in \mathrm{S}_N$ with $\sigma(i) = j$. For any local policy ν^i, define its relabeling for agent j by

$$(\sigma_* \nu^i)(b \mid s') := \nu^i(b \mid \sigma^{-1} \cdot s'). \tag{31}$$

This is a bijection between the local policy spaces of agents i and j. Let

$$\omega := (\nu^i, \pi^{-i}), \qquad \omega' := (\sigma_* \nu^i, \pi^{-j}).$$

For the deviating coordinate,

$$\omega'^{\sigma(i)}(b \mid \sigma \cdot s) = \omega^i(b|s)$$

by Eq. (31); for every $k \neq i$, the same identity holds between $\omega'^{\sigma(k)}$ and ω^k by permutation invariance of π. Thus the trajectory relabeling used in Lemma H.1, now between policies ω and ω', gives

$$V^j_{\omega'}(\sigma \cdot s) = V^i_{\omega}(s).$$

Averaging over the exchangeable initial distribution and reindexing states yields

$$V^j_{\sigma_* \nu^i, \pi^{-j}}(\mu) = V^i_{\nu^i, \pi^{-i}}(\mu).$$

Taking maxima and using the bijectivity of $\nu^i \mapsto \sigma_* \nu^i$ shows that the optimal response values are equal for agents i and j. Similarly, Eq. (24) and the exchangeability of μ imply

$$V^j_{\pi}(\mu) = V^i_{\pi}(\mu).$$

The best-response gains are therefore equal. □

Remark I.1 (Why the orbits must be pointed). Say that a state statistic read by a shared rule *certifies* the Nash gap if the unrestricted best response against invariant opponents can be written through it, so that driving the shared-rule gap to zero forces $\mathrm{NashGap}_M = 0$. The pointed statistic z^i certifies; coarser ones need not. Because z^i depends on which agent is asking, $\mathcal{Z}$ is not the state space of a smaller Markov *game*. For each fixed agent it is, however, the state space of that agent's best-response *MDP*: the stabilizer H_i has the z^i-fibers as its orbits, so Theorem 1 applied with H_i makes the inner maximum in Eq. (7) the optimal value of an MDP with $|\mathcal{Z}|$ states and $|\mathcal{B}|$ actions. Both the policy table and the gap evaluation are thus $|\mathcal{Z}|$-sized, not $|\mathcal{S}|$-sized, which is what makes the reduction computational rather than merely descriptive. Coarser statistics destroy this. Tying each rule to the agent's own local state alone, $\pi^i(\cdot \mid s) = \psi(\cdot \mid s^i)$, is still permutation invariant, but the best response, guaranteed only to be a function of z^i, need not be a function of s^i once an agent's reward or dynamics depend on the others. For a concrete failure take $N = 2$ with local states and actions in $\{0, 1\}$, dynamics that leave the state fixed, and $r^i(s, a) = \mathbf{1}[a^i = s^{-i}]$: this is agent-permutation symmetric, yet agent i's best response $a^i = s^{-i}$ reads the other agent's state. An optimum in the s^i-only class therefore leaves a profitable deviation and certifies nothing, whereas the pointed orbit z^i always does.

J Policy Parameterizations and Update Rules

This appendix records the two tabular parameterizations and the three learning dynamics that Appendix K substitutes quotient constants into, exactly as their sources state them, together with the shared variant used in Section 4.1. Abbreviations: PG is policy gradient (gradient play), NPG is natural policy gradient (natural gradient play). In all cases every agent updates its own parameters simultaneously and independently with a fixed step size $\eta > 0$, and the dynamics ascend agent i's own value V^i; the MPG identity of Definition 5 makes this the same as ascending the shared potential in the ith block, since $\nabla_{\vartheta^i} V^i_{\pi_\vartheta} = \nabla_{\vartheta^i} \Phi_{\pi_\vartheta}$.

Direct parameterization and projected PG (Zhang et al., 2024). Each local policy is a probability vector $\pi^i(\cdot \mid s) \in \Delta(\mathcal{A}^i)$ and the parameters ϑ are the table entries themselves, constrained to $\mathcal{X} := \prod_{i,s} \Delta(\mathcal{A}^i)$. Gradient play takes a step in those entries and restores feasibility by Euclidean projection,

$$\vartheta_{t+1} = \mathrm{Proj}_{\mathcal{X}}\big(\vartheta_t + \eta \, \nabla_\vartheta \Phi_{\pi_{\vartheta_t}}(\mu)\big), \qquad \eta = \frac{(1-\gamma)^3}{2 \sum_i |\mathcal{A}^i|}.$$

No regularizer is used. Zhang et al. (2024) prove that the average of $\mathrm{NashGap}_M(\pi_{\vartheta_t})^2$ over the T iterates produced is at most ϵ^2, hence that some iterate satisfies $\mathrm{NashGap}_M(\pi_{\vartheta_t}) \leq \epsilon$, whenever $T \geq 64 \kappa^2 \Delta_\Phi \, |\mathcal{S}| \sum_i \left|\mathcal{A}^i\right| / [(1-\gamma)^3 \epsilon^2]$. This matches the min-over-iterates form of Corollary 1, whose window $0 \leq t < T$ indexes those same T iterates; the three sources differ from one another only in this labeling.

Softmax parameterization. A logit table $\vartheta = (\vartheta^i_{s,b})_{i,s,b}$ induces

$$\pi^i_\vartheta(b|s) = \frac{\exp(\vartheta^i_{s,b})}{\sum_{b' \in \mathcal{A}^i} \exp(\vartheta^i_{s,b'})}. \tag{32}$$

Both softmax dynamics below add a *log-barrier* regularizer: a multiple $\lambda > 0$ of a sum of $\log \pi^i_\vartheta$ terms. It diverges to $-\infty$ at the simplex boundary, so ascending it repels the iterates from faces where a probability vanishes; this is what makes an approximate first-order stationary point provably near-Nash, and it is the multi-agent form of the single-agent barrier of Agarwal et al. (2021). The two sources normalize it differently, and we keep each source's own form.

Log-barrier softmax PG (Chen et al., 2022b). The regularized objective sums one barrier per agent, each averaged over that agent's own logits, and is stated so that the added term is a negative average Kullback–Leibler divergence to the uniform action distribution:

$$L_\lambda(\vartheta) := \Phi_{\pi_\vartheta}(\mu) - \lambda \sum_{i=1}^N \mathbb{E}_{s\sim \mathrm{Unif}_\mathcal{S}} \left[\mathrm{KL}\big(\mathrm{Unif}_{\mathcal{A}^i} \,\big\|\, \pi^i_\vartheta(\cdot \mid s)\big)\right] \tag{33}$$

$$= \Phi_{\pi_\vartheta}(\mu) + \lambda \sum_{i=1}^N \left[\frac{\sum_{s,b} \log \pi^i_\vartheta(b|s)}{|\mathcal{S}|\,|\mathcal{A}^i|} + \log \left|\mathcal{A}^i\right| \right]. \tag{34}$$

The dynamics are plain gradient ascent $\vartheta_{t+1} = \vartheta_t + \eta \nabla_\vartheta L_\lambda(\vartheta_t)$ from $\vartheta_0 = 0$ with $\lambda = \epsilon/(2\kappa)$ and $\eta = 1/\beta_\lambda$, where $\beta_\lambda := 41N/[4(1-\gamma)^3] + 2\lambda N/|\mathcal{S}|$ bounds the smoothness of L_λ: the first summand is the smoothness of Φ in the logits and the second is N times the per-agent barrier smoothness $2\lambda/|\mathcal{S}|$. Their key lemma is that $\|\nabla_\vartheta L_\lambda(\vartheta)\|_2 \le \lambda/(2\,|\mathcal{S}| \max_i |\mathcal{A}^i|)$ already implies a $2\lambda\kappa$-Nash policy; their iteration bound is the first two-term expression of Corollary K.1, with leading constants 328 and $32N$.

Log-barrier softmax NPG (Zhang et al., 2022). Here the barrier is an unnormalized sum, deliberately independent of the visitation measure so that agent i's gradient depends only on ϑ^i and the dynamics stay decentralized:

$$\widetilde{\Phi}(\vartheta) := \Phi_{\pi_\vartheta}(\mu) + \lambda \sum_{i=1}^N \sum_{s,b} \log \pi^i_\vartheta(b|s).$$

Natural gradient play preconditions by the Moore–Penrose pseudoinverse of agent i's Fisher information matrix F^i,

$$\vartheta^i_{t+1} = \vartheta^i_t + \eta\, F^i(\vartheta_t)^\dagger \nabla_{\vartheta^i} \widetilde{\Phi}(\vartheta_t),$$

which for the softmax table is equivalent to the multiplicative form $\pi^i_{t+1}(b|s) \propto \pi^i_t(b|s) \exp\big(\frac{\eta}{1-\gamma} \mathring{A}^i_t(s,b) + \frac{\eta\lambda}{d_t(s)\pi^i_t(b|s)} - \frac{\eta\lambda|\mathcal{A}^i|}{d_t(s)}\big)$, with d_t the normalized visitation. The corresponding iteration bound is the third line of Corollary K.1.

Remark J.1 (Reading the constants across sources). Three conventions differ between these papers, and Corollary K.1 is stated in ours. First, visitation: Zhang et al. (2024) and Chen et al. (2022b) use the unnormalized d^π_μ of Eq. (13), whereas Zhang et al. (2022) normalize by $1-\gamma$. Second, the symbol M: in Zhang et al. (2024) and Chen et al. (2022b) it is the mismatch *ratio* $\max_{\pi,\pi'} \left\| d^\pi_\mu / d^{\pi'}_\mu \right\|_\infty$, our κ, while in Zhang et al. (2022) it is the *coverage* quantity $\sup_\pi \max_s 1/d^\pi_\mu(s)$ formed from their normalized visitation, which is exactly the coverage constant C_{cov} defined in Appendix K; those authors state the distinction from the mismatch ratio explicitly. This is why the first two lines of Corollary K.1 carry κ_G and the third carries $C_{\mathrm{cov},G}$. Third, potential scale: the first two bounds involve the total-potential range Δ_Φ, the third the range Δ_ϕ of the potential function.

Shared softmax with a log-barrier. For an agent-permutation-symmetric X, tie the softmax Eq. (32) across the pointed orbits: one table $\vartheta = (\vartheta_{z,b})_{z\in\mathcal{Z}_X,\, b\in\mathcal{B}_X}$ gives

$$\psi_\vartheta(b \mid z) = \frac{\exp(\vartheta_{z,b})}{\sum_{b'\in\mathcal{B}_X} \exp(\vartheta_{z,b'})}, \qquad \pi^i_\vartheta(b \mid x) := \psi_\vartheta(b \mid z^i_X(x)).$$

The objective is Eq. (34) adapted to this single tied table, carrying *one* barrier term over the $|\mathcal{Z}_X|$ pointed orbits rather than one per agent:

$$\widehat{\mathcal{L}}^X_\lambda(\vartheta) := \Phi^X_{\pi_\vartheta}(\mu_X) + \lambda \left[\frac{1}{|\mathcal{Z}_X|\,|\mathcal{B}_X|} \sum_{z\in\mathcal{Z}_X} \sum_{b\in\mathcal{B}_X} \log \psi_\vartheta(b \mid z) + \log|\mathcal{B}_X| \right], \tag{35}$$

whose bracket is $-\frac{1}{|\mathcal{Z}_X|}\sum_{z\in\mathcal{Z}_X} \mathrm{KL}\big(\mathrm{Unif}_{\mathcal{B}_X} \,\big\|\, \psi_\vartheta(\cdot \mid z)\big) \le 0$, vanishing at $\vartheta = 0$. Corollary 1 runs gradient ascent $\vartheta_{t+1} = \vartheta_t + \eta\,\nabla\widehat{\mathcal{L}}^X_\lambda(\vartheta_t)$ from $\vartheta_0 = 0$ with $\lambda = \epsilon/(2\kappa_X)$ and $\eta = \big(z_{\max,X}\,41N/[4(1-\gamma)^3] + \lambda/|\mathcal{Z}_X|\big)^{-1}$. Two consequences of the single tied barrier are visible in the rate of Corollary 1: its smoothness contribution is $\lambda/|\mathcal{Z}_X|$ rather than $2\lambda N/|\mathcal{S}|$, which is why the lower-order term carries neither a factor N nor a factor 2, and the leading term instead acquires the tiling multiplicity $z_{\max,X}$.

K Quotient Constants and Existing Bounds

This appendix collects the game-dependent constants used by the existing MPG bounds, records how each behaves under state quotienting, and states the three substitutions summarized in Section 4.1. Throughout, M is a G-state-symmetric MPG carrying the G-invariant potential ϕ_G of Lemma C.1, which we continue to denote by ϕ, and G acts freely in the sense of Section 4.1, so that every orbit has exactly $|G|$ elements and $|\overline{\mathcal{S}}| = |\mathcal{S}|/|G|$.

The total-potential range $\Delta_{\Phi,X}$ and the mismatch coefficient κ_X of a finite MPG X are those of Section 4.1, and $\Phi_{\sup,X}$ is $\Phi_{\sup}$ of Section 4.2 evaluated on X; writing Φ^X_π for the total potential of a game X and $\Phi_{\max,X}$, $\Phi_{\min,X}$ for its extrema over product policies and states, which exist by finiteness,

$$\Delta_{\Phi,X} = \Phi_{\max,X} - \Phi_{\min,X}, \qquad \Phi_{\sup,X} = \max\{|\Phi_{\min,X}|, |\Phi_{\max,X}|\}. \tag{36}$$

One constant is new here: when $d^\pi_{\mu_X}(x) > 0$ throughout, the coverage constant is $C_{\mathrm{cov},X} := \max_{\pi\in\Pi(X),x}[(1-\gamma)d^\pi_{\mu_X}(x)]^{-1}$. Abbreviate $\Delta_\Phi := \Delta_{\Phi,M}$, $\kappa := \kappa_M$ and $C_{\mathrm{cov}} := C_{\mathrm{cov},M}$, note that $\Phi_{\sup} = \Phi_{\sup,M}$, and write $\Delta_{\Phi,G}$, $\Phi_{\sup,G}$, κ_G, $C_{\mathrm{cov},G}$ for the corresponding quotient quantities, $A_{\max} := \max_i |\mathcal{A}^i|$, and $\Delta_\phi := \max_{s,a}\phi(s,a) - \min_{s,a}\phi(s,a)$ for the range of the potential function, which equals that of $\overline{\phi}$.

By Lemma D.1, $\Delta_{\Phi,G} \le \Delta_\Phi$, $\Phi_{\sup,G} \le \Phi_{\sup}$, and $\kappa_G \le \kappa$; positive visitation is inherited, so κ_G and $C_{\mathrm{cov},G}$ are finite whenever κ and C_{cov} are. The coverage constant behaves as follows.

Lemma K.1 (Quotient coverage)**.** *Assume the positive-visitation hypothesis of Lemma D.1(b). Then $C_{\mathrm{cov},G} \le C_{\mathrm{cov}}$. If in addition μ is G invariant and G acts freely, then $C_{\mathrm{cov},G} \le C_{\mathrm{cov}}/|G|$.*

Proof. By Eq. (16), $\overline{d}^{\overline{\pi}}_{\overline{\mu}}(\bar{s}) = \sum_{s\in\bar{s}} d^{\overline{\pi}^\uparrow}_\mu(s) \ge d^{\overline{\pi}^\uparrow}_\mu(s)$ for every $s \in \bar{s}$, and lifting maps $\Pi(\overline{M})$ into $\Pi(M)$, which gives the first claim. Suppose now that μ is G invariant, and let π be G invariant. Policy invariance and Eq. (2) give $P^\pi(L_g[s'] \mid L_g[s]) = P^\pi(s' \mid s)$ for the induced state kernel, hence $p^\pi_t(L_g[s'] \mid L_g[s]) = p^\pi_t(s' \mid s)$ for every $t \ge 0$ by induction. Reindexing the initial state as $s_0 = L_g[u]$ and using $\mu(L_g[u]) = \mu(u)$ therefore yields $d^\pi_\mu(L_g[s]) = d^\pi_\mu(s)$, so the lifted occupancy is constant on orbits; if moreover G acts freely, each orbit has $|G|$ elements and $\overline{d}^{\overline{\pi}}_{\overline{\mu}}(\bar{s}) = |G|\; d^{\overline{\pi}^\uparrow}_\mu(s)$, whence the second claim. □

Corollary K.1 (Representative quotient optimization rates)**.** *Assume positive visitation, as the mismatch-based source bounds require. Each bound below, run on $\overline{M}$ under its own hypotheses, initialization, regularization, and step size, with every game-dependent constant evaluated on $\overline{M}$, gives the stated sufficient iteration count T_{quot} and quotient/full ratio. Parenthesized values are the resulting bounds when G acts freely.*

- Gradient play, direct parameterization *(projected PG) (Zhang et al., 2024, Theorem 3):* $T_{\mathrm{quot}} = O\big(\kappa_G^2 |\overline{\mathcal{S}}|\,\Delta_{\Phi,G}\sum_i |\mathcal{A}^i| / [(1-\gamma)^3\epsilon^2]\big)$*; ratio at most* $\kappa_G^2|\overline{\mathcal{S}}|/(\kappa^2|\mathcal{S}|) \le |\overline{\mathcal{S}}|/|\mathcal{S}|$ *(*$|G|^{-1}$*).*
- Log-barrier softmax PG *(Chen et al., 2022b, Theorem 3.6):* $T_{\mathrm{quot}} = O\big(NA^2_{\max}\Delta_{\Phi,G}[\kappa_G^2|\overline{\mathcal{S}}|^2/((1-\gamma)^3\epsilon^2) + \kappa_G|\overline{\mathcal{S}}|/\epsilon]\big)$*; termwise ratios* $(\kappa_G|\overline{\mathcal{S}}|/(\kappa|\mathcal{S}|))^2$ *and* $\kappa_G|\overline{\mathcal{S}}|/(\kappa|\mathcal{S}|)$ *(*$|G|^{-2}, |G|^{-1}$*).*

- Log-barrier natural gradient play *(softmax NPG) (Zhang et al., 2022, Theorem 7 and Table 1):* $T_{\mathrm{quot}} = O\big(NA_{\max}\Delta_\phi C^2_{\mathrm{cov},G}/[(1-\gamma)^4\epsilon^2]\big)$*; ratio* $(C_{\mathrm{cov},G}/C_{\mathrm{cov}})^2 \le 1$*, and* $|G|^{-2}$ *when* μ *is* G *invariant and* G *acts freely.*

Each line is a direct substitution: the parameterizations are those of Appendix J, the reduced game $\overline{M}$ is an MPG by Proposition 1, its rewards still lie in $[0,1]$, its action sets are unchanged, its positive visitation is inherited from M by Lemma D.1(b), and each cited theorem's remaining hypotheses concern only the parameterization and step size and so are satisfied verbatim on $\overline{M}$. Two caveats. The displayed figures are ratios of *sufficient upper bounds*, not measured runtimes, and the omitted factor $\Delta_{\Phi,G}/\Delta_\Phi \le 1$ can only improve the first two comparisons. The reduction changes the constants a bound is evaluated at, nothing more.

L Proof of the Shared Log-Barrier Rate

Corollary 1 is stated for M. We prove it for a general finite agent-permutation-symmetric MPG X with positive visitation, writing $\mathcal{S}_X$ and $\mathcal{B}_X$ for its state space and local action set, $\mathcal{Z}_X := (\mathcal{N}\times\mathcal{S}_X)/\mathrm{S}_N$ and $z^i_X(x) := [(i,x)]_{\mathrm{S}_N}$ for its pointed orbits, $z_{\max,X} := \max_z \big|\{(i,x) : z^i_X(x) = z\}\big|$, and κ_X, $\Delta_{\Phi,X}$ for its mismatch and potential range, so that the case $X = M$ is the corollary and Lemma L.1 below can also take $X = \overline{M}$.

Proof of Corollary 1. Abbreviate $\kappa = \kappa_X$ and $z_{\max} = z_{\max,X}$ within this proof. For a shared policy π_ϑ, define its orbit visitation by

$$\mathfrak{d}_\vartheta(z) := \sum_{i=1}^{N}\sum_{x:\, z^i_X(x)=z} d^{\pi_\vartheta}_{\mu_X}(x).$$

It is positive because every pointed orbit has a nonempty fiber and X satisfies positive visitation.

Step 1: shared policy gradient. Permutation equivariance of values and marginalized local advantages, proved by the same relabeling argument as Lemma H.1, implies that

$$z^i_X(x) = z^j_X(y) \quad\Longrightarrow\quad \mathring{A}^i_{\pi_\vartheta}(x,b) = \mathring{A}^j_{\pi_\vartheta}(y,b).$$

Let $\widehat{\mathring{A}}_\vartheta(z,b)$ denote this common value.

Two facts turn this into a gradient formula. First, the MPG identity of Definition 5 says that $\Phi_\pi(x) - V^i_\pi(x)$ is unchanged when only agent i's policy varies, so it depends on π^{-i} alone; differentiating in agent i's own logits therefore gives

$$\frac{\partial\Phi^X_{\pi_\vartheta}(\mu_X)}{\partial\vartheta^i_{x,b}} = \frac{\partial V^i_{\pi_\vartheta}(\mu_X)}{\partial\vartheta^i_{x,b}} = d^{\pi_\vartheta}_{\mu_X}(x)\,\pi^i_\vartheta(b\mid x)\,\mathring{A}^i_{\pi_\vartheta}(x,b), \tag{37}$$

the second equality being the tabular softmax policy-gradient identity (Agarwal et al., 2021) applied to agent i's best-response MDP against π^{-i}_ϑ, whose advantage is exactly $\mathring{A}^i_{\pi_\vartheta}$; the factor $(1-\gamma)^{-1}$ of the usual statement is absent because d^π_μ is unnormalized here. Second, the shared logit $\vartheta_{z,b}$ enters every coordinate $\vartheta^i_{x,b}$ with $z^X_i(x) = z$, so the chain rule gives

$$\begin{aligned}\frac{\partial\Phi^X_{\pi_\vartheta}(\mu_X)}{\partial\vartheta_{z,b}} &= \sum_{i=1}^{N}\sum_{x:\, z^i_X(x)=z} d^{\pi_\vartheta}_{\mu_X}(x)\pi^i_\vartheta(b\mid x)\mathring{A}^i_{\pi_\vartheta}(x,b)\\ &= \mathfrak{d}_\vartheta(z)\psi_\vartheta(b\mid z)\widehat{\mathring{A}}_\vartheta(z,b).\end{aligned}$$

Differentiating the orbit-uniform barrier in Eq. (35) therefore yields

$$\frac{\partial\widehat{\mathcal{L}}^X_\lambda(\vartheta)}{\partial\vartheta_{z,b}} = \mathfrak{d}_\vartheta(z)\psi_\vartheta(b\mid z)\widehat{\mathring{A}}_\vartheta(z,b) + \frac{\lambda}{|\mathcal{Z}_X|}\left(\frac{1}{|\mathcal{B}_X|} - \psi_\vartheta(b\mid z)\right). \tag{38}$$

Step 2: approximate stationarity controls unrestricted deviations. Suppose

$$\left\|\nabla \widehat{\mathcal{L}}_\lambda^X(\vartheta)\right\|_2 \le \frac{\lambda}{2\,|\mathcal{Z}_X|\,|\mathcal{B}_X|}. \tag{39}$$

Fix (z,b) with $\widehat{\mathring{A}}_\vartheta(z,b) \ge 0$. The first term on the right of Eq. (38) is then nonnegative. Combining this with the componentwise upper bound implied by Eq. (39) gives

$$\frac{\lambda}{2\,|\mathcal{Z}_X|\,|\mathcal{B}_X|} \ge \frac{\lambda}{|\mathcal{Z}_X|}\left(\frac{1}{|\mathcal{B}_X|} - \psi_\vartheta(b \mid z)\right), \qquad \text{hence} \qquad \psi_\vartheta(b \mid z) \ge \frac{1}{2\,|\mathcal{B}_X|}.$$

Solving Eq. (38) for the advantage, using the componentwise gradient bound and $\psi_\vartheta(b \mid z)^{-1} \le 2\,|\mathcal{B}_X|$, gives

$$\begin{aligned}\widehat{\mathring{A}}_\vartheta(z,b) &= \frac{1}{\mathfrak{d}_\vartheta(z)}\left[\frac{1}{\psi_\vartheta(b \mid z)}\frac{\partial \widehat{\mathcal{L}}_\lambda^X}{\partial \vartheta_{z,b}} + \frac{\lambda}{|\mathcal{Z}_X|}\left(1 - \frac{1}{|\mathcal{B}_X|\,\psi_\vartheta(b \mid z)}\right)\right] \\ &\le \frac{2\lambda}{|\mathcal{Z}_X|\,\mathfrak{d}_\vartheta(z)}.\end{aligned} \tag{40}$$

The same upper bound is automatic when the advantage is negative.

Fix any agent i and an unrestricted stationary local deviation ν^i, and write $\pi' = (\nu^i, \pi_\vartheta^{-i})$. The performance-difference identity and Eq. (40) imply

$$\begin{aligned}V^i_{\nu^i,\pi_\vartheta^{-i}}(\mu_X) - V^i_{\pi_\vartheta}(\mu_X) &= \sum_x d^{\pi'}_{\mu_X}(x) \sum_b \nu^i(b \mid x)\mathring{A}^i_{\pi_\vartheta}(x,b) \\ &\le \frac{2\lambda}{|\mathcal{Z}_X|}\sum_x \frac{d^{\pi'}_{\mu_X}(x)}{\mathfrak{d}_\vartheta(z^i_X(x))}.\end{aligned}$$

For each pointed orbit z, mismatch gives

$$\sum_{x:\, z^i_X(x)=z} d^{\pi'}_{\mu_X}(x) \le \kappa \sum_{x:\, z^i_X(x)=z} d^{\pi_\vartheta}_{\mu_X}(x) \le \kappa\, \mathfrak{d}_\vartheta(z).$$

Grouping by pointed orbit shows that the final sum is at most $\kappa\,|\mathcal{Z}_X|$. Therefore every unrestricted deviation satisfies

$$V^i_{\nu^i,\pi_\vartheta^{-i}}(\mu_X) - V^i_{\pi_\vartheta}(\mu_X) \le 2\lambda\kappa. \tag{41}$$

Maximizing over ν^i and i proves that Eq. (39) implies $\mathrm{NashGap}_X(\pi_\vartheta) \le 2\lambda\kappa$.

Step 3: reaching the stationarity threshold. Let T be the linear tiling map from shared logits to full-table logits, written T rather than Θ to avoid collision with asymptotic notation:

$$(\mathsf{T}\vartheta)^i_{x,b} = \vartheta_{z^i_X(x),b}.$$

Its Gram matrix is diagonal, with $n_X(z) := \big|\{(i,x) : z^X_i(x) = z\}\big|$:

$$(\mathsf{T}^\top \mathsf{T}\,\vartheta)_{z,b} = n_X(z)\vartheta_{z,b}.$$

Consequently, $\|\mathsf{T}\|_2^2 = z_{\max}$. The full-logit potential smoothness used in Chen et al. (2022b) is $41N/[4(1-\gamma)^3]$ (Appendix J). That constant needs no hypothesis on ϕ_X: because $\nabla_{\vartheta^j}\nabla_{\vartheta^i}\Phi^X = \nabla_{\vartheta^j}\nabla_{\vartheta^i}V^i$ for every pair (i,j) by Eq. (37), every Hessian block of Φ^X is a block of some individual value function, and it is bounded using only $r^i \in [0,1]$ together with the block-matrix argument of Leonardos et al. (2022) and the softmax derivative bounds of Agarwal et al. (2021). The chain rule therefore gives shared-coordinate potential smoothness

$$z_{\max}\frac{41N}{4(1-\gamma)^3}.$$

The Hessian of the barrier is block diagonal over z; each block is $-\lambda/|\mathcal{Z}_X|$ times the covariance matrix of a categorical distribution, whose spectral norm is at most one. Thus

$$\widehat{\beta}_\lambda := z_{\max}\frac{41N}{4(1-\gamma)^3} + \frac{\lambda}{|\mathcal{Z}_X|} \tag{42}$$

is a smoothness constant for $\widehat{\mathcal{L}}^X_\lambda$.

Gradient ascent with step $1/\widehat{\beta}_\lambda$ satisfies

$$\min_{0\le t<T}\left\|\nabla\widehat{\mathcal{L}}^X_\lambda(\vartheta_t)\right\|_2^2 \le \frac{2\widehat{\beta}_\lambda\Delta_{\Phi,X}}{T}.$$

Indeed, the barrier is nonpositive because it is the negative average $\mathrm{KL}(\mathrm{Unif}_{\mathcal{B}_X}\|\psi_\vartheta(\cdot\mid z))$, it vanishes at $\vartheta_0=0$, and hence the objective increase from initialization is at most $\Delta_{\Phi,X}$. It follows that some iterate satisfies Eq. (39) whenever

$$T \ge \frac{8\widehat{\beta}_\lambda\,|\mathcal{Z}_X|^2\,|\mathcal{B}_X|^2\,\Delta_{\Phi,X}}{\lambda^2}.$$

Substitute $\lambda=\epsilon/(2\kappa)$ and Eq. (42) to obtain

$$T \ge \frac{328 z_{\max}N\kappa^2\,|\mathcal{Z}_X|^2\,|\mathcal{B}_X|^2\,\Delta_{\Phi,X}}{(1-\gamma)^3\epsilon^2} + \frac{16\kappa DB^2\Delta_{\Phi,X}}{\epsilon}.$$

At that iterate, Eq. (41) is at most ϵ; the leading (order ϵ^{-2}) term is the rate in Corollary 1. □

The comparison in Section 4.1 between the shared leading term and its full-table counterpart turns on $z_{\max,X}\,|\mathcal{Z}_X|^2$ versus $|\mathcal{S}_X|^2$. The two claims made there are the following.

Lemma L.1 (Tiling multiplicity and when sharing helps)**.** *Let X be a finite agent-permutation-symmetric Markov game, with $z_{\max,X}$ as above.*

(a) In the canonical model $\mathcal{S}_X=\mathcal{C}^N$ with the coordinate action,

$$z_{\max,X} \le N\,|\mathcal{C}|^{N-1}, \qquad |\mathcal{Z}_X| = |\mathcal{C}|\binom{|\mathcal{C}|+N-2}{N-1} = O\Big(N^{|\mathcal{C}|-1}\Big).$$

Hence, for fixed $|\mathcal{C}|\ge 2$, $z_{\max,X}\,|\mathcal{Z}_X|^2 = O\Big(N^{2|\mathcal{C}|-1}\,|\mathcal{C}|^{N-1}\Big)$ while $|\mathcal{S}_X|^2=|\mathcal{C}|^{2N}$, so $z_{\max,X}\,|\mathcal{Z}_X|^2<|\mathcal{S}_X|^2$ for all sufficiently large N.

(b) Let $X=\overline{M}$ for the ring of Section 3.3 with m sites and $N\ge 2$ agents, and write

$$\mathcal{M}_m(k) := \max\left\{\frac{k!}{m_0!\cdots m_{m-1}!}\ :\ m_c\ge 0,\ \sum_c m_c = k\right\}.$$

Then $|\mathcal{S}_X|=m^{N-1}$, $|\mathcal{Z}_X|=\binom{m+N-2}{N-1}$, and $z_{\max,X}=N\,\mathcal{M}_m(N-1)$. At $m=6$ the ratio $\rho(N):=z_{\max,X}\,|\mathcal{Z}_X|^2/\,|\mathcal{S}_X|^2$ exceeds one for $2\le N\le 5$, is strictly decreasing for $N\ge 5$, and satisfies $\rho(6)<1$. So $z_{\max,X}\,|\mathcal{Z}_X|^2<|\mathcal{S}_X|^2$ fails exactly for $2\le N\le 5$ and holds for every $N\ge 6$.

Proof. Part (a). Fix $z=(c,\{\{c_2,\dots,c_N\}\})\in\mathcal{Z}_X$. A pair (i,x) with $z_i^X(x)=z$ is determined by the focal index i, of which there are N choices, together with the remaining $N-1$ coordinates of x, of which there are at most $|\mathcal{C}|^{N-1}$ assignments in total. Hence $z_{\max,X}\le N\,|\mathcal{C}|^{N-1}$. The count $|\mathcal{Z}_X|$ is Lemma G.1, and

$$\binom{|\mathcal{C}|+N-2}{N-1} = \binom{N+|\mathcal{C}|-2}{|\mathcal{C}|-1} \le \frac{(N+|\mathcal{C}|-2)^{|\mathcal{C}|-1}}{(|\mathcal{C}|-1)!},$$

which is $O(N^{|\mathcal{C}|-1})$ at fixed $|\mathcal{C}|$. Multiplying the two estimates gives the stated bound on $z_{\max,X}\,|\mathcal{Z}_X|^2$, and

$$\frac{|\mathcal{C}|^{2N}}{N^{2|\mathcal{C}|-1}\,|\mathcal{C}|^{N-1}} = \frac{|\mathcal{C}|^{N+1}}{N^{2|\mathcal{C}|-1}} \longrightarrow \infty \qquad (N \to \infty),$$

since $|\mathcal{C}| \geq 2$ makes the numerator exponential and the denominator polynomial in N.

Part (b). Here $\mathcal{S} = C_m^N$ and $G = C_m$ acts freely by global rotation, so $\left|\overline{\mathcal{S}}\right| = m^{N-1}$. By Lemma M.1, $|\mathcal{Z}_X| = |\mathcal{Z}_{\overline{M}}| = |\mathcal{W}|$, which is the number of multisets of size $N-1$ drawn from the m focal-relative offsets, namely $\binom{m+N-2}{N-1}$.

For the multiplicity, S_N acts on $\mathcal{N} \times \overline{\mathcal{S}}$ and the fiber of z under $(i, \bar{s}) \mapsto z_i^{\overline{M}}(\bar{s})$ is a single orbit, so orbit–stabilizer gives $n_X(z) = N!/\,|\mathrm{Stab}(i, \bar{s})|$ for any $(i, \bar{s})$ in that fiber. If $\sigma(i) = i$ and $\rho(\sigma \cdot s) = \rho(s)$, then $\sigma \cdot s = L_g[s]$ for some $g \in C_m$, and comparing ith coordinates gives $s^i = (\sigma \cdot s)^i = (L_g[s])^i = s^i + g$, so $g = 0$. Hence

$$\mathrm{Stab}(i, \bar{s}) = \{\sigma \in \mathrm{S}_N : \sigma(i) = i,\ \sigma \cdot s = s\},$$

the permutations of $\mathcal{N} \setminus \{i\}$ preserving the coordinates of s, whose order is $\prod_c m_c!$ for the multiplicity vector (m_c) of the offset multiset $z = \{\{s^j - s^i : j \neq i\}\}$. Therefore

$$n_X(z) = \frac{N!}{\prod_c m_c!} = N\,\frac{(N-1)!}{\prod_c m_c!},$$

and maximizing over multiplicity vectors of total size $N-1$ gives $z_{\max,X} = N\,\mathcal{M}_m(N-1)$.

For the monotonicity, first note $\mathcal{M}_m(k) \leq m\,\mathcal{M}_m(k-1)$ for $k \geq 1$: if (m_c) attains $\mathcal{M}_m(k)$, some $c^\star$ has $m_{c^\star} \geq k/m$, and writing $m'_c := m_c - \mathbf{1}\{c = c^\star\}$,

$$\frac{k!}{\prod_c m_c!} = \frac{k}{m_{c^\star}}\,\frac{(k-1)!}{\prod_c m'_c!} \leq m\,\mathcal{M}_m(k-1).$$

Since $\binom{m+N-1}{N}\big/\binom{m+N-2}{N-1} = (N+m-1)/N$, and $|\mathcal{S}_X|^2$ is multiplied by m^2 when N increases by one,

$$\frac{\rho(N+1)}{\rho(N)} \leq \frac{(N+1)m}{N} \cdot \frac{(N+m-1)^2}{N^2} \cdot \frac{1}{m^2} = \frac{(N+1)(N+m-1)^2}{mN^3}.$$

At $m = 6$ the right-hand side is below one for every $N \geq 5$, because

$$6N^3 - (N+1)(N+5)^2 = 5N^3 - 11N^2 - 35N - 25$$

equals 150 at $N = 5$ and has derivative $15N^2 - 22N - 35$, which is positive at $N = 5$ and increasing thereafter. Thus ρ is strictly decreasing on $N \geq 5$. Finally, $\mathcal{M}_6(k) = k!$ for $k \leq 6$, since then the items may be placed in distinct bins, so direct evaluation gives

$$\rho(2) = \frac{2 \cdot 6^2}{6^2} = 2, \qquad \rho(3) = \frac{6 \cdot 21^2}{36^2} = \frac{2646}{1296}, \qquad \rho(4) = \frac{24 \cdot 56^2}{216^2} = \frac{75264}{46656},$$
$$\rho(5) = \frac{120 \cdot 126^2}{1296^2} = \frac{1905120}{1679616}, \quad \rho(6) = \frac{720 \cdot 252^2}{7776^2} = \frac{45722880}{60466176}.$$

The first four exceed one and $\rho(6) < 1$; combined with the strict decrease from $N = 5$ onward, this gives the dichotomy. □

M Proof of the Combined Reduction

We first record the orbit structure that the theorem rests on and that justifies composing the two reductions in either order.

Lemma M.1 (Orbit structure under commuting symmetries)**.** *Let M have commuting symmetries (Definition 4).*

(a) Agent permutations induce a well-defined action on $\overline{M}$ by $\sigma \cdot \bar{s} := \rho(\sigma \cdot s)$, under which $\overline{M}$ is agent-permutation-symmetric. If μ is exchangeable, then $\overline{\mu}$ is permutation invariant, and if M is an MPG, then the quotient potential $\overline{\phi}$ built from a $G \times \mathrm{S}_N$-invariant potential is permutation invariant.

(b) The two reduction orders agree,

$$\mathcal{W} \;\cong\; (\mathcal{N} \times \overline{\mathcal{S}})/\mathrm{S}_N \;\cong\; \mathcal{Z}/G, \tag{43}$$

where each $\cong$ is a canonical bijection of orbit spaces, so $\mathcal{W}$ is reached by quotienting by G then sharing or by sharing then quotienting by G.

(c) The product policies carried by one rule $\psi : \mathcal{W} \to \Delta(\mathcal{B})$ through $\pi^i(b|s) = \psi(b|w^i(s))$ are exactly the jointly invariant ones.

Proof. **Part (a): the descended permutation action.** Suppose $s, \widetilde{s} \in \bar{s}$, so $\widetilde{s} = L_g[s]$ for some $g \in G$. By Eq. (8),

$$\sigma \cdot \widetilde{s} = \sigma \cdot L_g[s] = L_g[\sigma \cdot s].$$

Thus $\sigma \cdot \widetilde{s}$ and $\sigma \cdot s$ belong to the same G orbit, so the descended action $\sigma \cdot \bar{s} = \rho(\sigma \cdot s)$ is independent of the representative. It is a group action because the original S_N action is one.

For quotient states $\bar{s}, \bar{s}'$ with representatives s, s' and any joint action a, reindex the target orbit by $u = \sigma \cdot v$ and use Eq. (5):

$$\begin{aligned}
\overline{P}(\sigma \cdot \bar{s}' \mid \sigma \cdot \bar{s}, \sigma \cdot a) &= \sum_{u \in \sigma \cdot \bar{s}'} P(u \mid \sigma \cdot s, \sigma \cdot a) \\
&= \sum_{v \in \bar{s}'} P(\sigma \cdot v \mid \sigma \cdot s, \sigma \cdot a) \\
&= \sum_{v \in \bar{s}'} P(v \mid s, a) = \overline{P}(\bar{s}' \mid \bar{s}, a).
\end{aligned}$$

The reward identity follows directly:

$$\overline{r}^{\,\sigma(i)}(\sigma \cdot \bar{s}, \sigma \cdot a) = r^{\sigma(i)}(\sigma \cdot s, \sigma \cdot a) = r^i(s, a) = \overline{r}^{\,i}(\bar{s}, a).$$

Thus $\overline{M}$ is agent-permutation-symmetric. If μ is exchangeable, reindexing the orbit sum gives

$$\overline{\mu}(\sigma \cdot \bar{s}) = \sum_{u \in \sigma \cdot \bar{s}} \mu(u) = \sum_{v \in \bar{s}} \mu(\sigma \cdot v) = \overline{\mu}(\bar{s}).$$

If M is an MPG, choose the $G \times \mathrm{S}_N$-invariant potential from Lemma C.1. Then

$$\overline{\phi}(\sigma \cdot \bar{s}, \sigma \cdot a) = \phi(\sigma \cdot s, \sigma \cdot a) = \phi(s, a) = \overline{\phi}(\bar{s}, a),$$

so the quotient potential is permutation invariant.

Part (b): the two orbit orders agree. Commutativity makes the displayed formula in Eq. (9) a left action of $G \times \mathrm{S}_N$. Define

$$\begin{aligned}
F : \mathcal{W} &\longrightarrow \big(\mathcal{N} \times (\mathcal{S}/G)\big)/\mathrm{S}_N, \\
F([(i, s)]_{G \times \mathrm{S}_N}) &:= [(i, \rho(s))]_{\mathrm{S}_N}.
\end{aligned}$$

The representative calculation above shows that F is well defined. It is surjective because every pointed quotient state has a representative in $\mathcal{N} \times \mathcal{S}$. If $F([(i, s)]) = F([(j, s')])$, there is a $\sigma \in \mathrm{S}_N$ with $j = \sigma(i)$ and $\rho(s') = \rho(\sigma \cdot s)$. Therefore $s' = L_g[\sigma \cdot s]$ for some $g \in G$, so the two pointed states are in the same $G \times \mathrm{S}_N$ orbit. Thus F is injective.

Similarly, commutativity induces a well-defined G action on $\mathcal{Z} = (\mathcal{N} \times \mathcal{S})/\mathrm{S}_N$ by

$$g \cdot [(i, s)]_{\mathrm{S}_N} := [(i, L_g[s])]_{\mathrm{S}_N}.$$

Quotienting this orbit space by G identifies exactly the same pairs of pointed states as the product action. This proves both isomorphisms in Eq. (43).

Part (c): jointly invariant policies. If π is jointly invariant and $(j, s') = (\sigma(i), L_g[\sigma \cdot s])$, then

$$\begin{aligned}\pi^j(b \mid s') &= \pi^{\sigma(i)}(b \mid L_g[\sigma \cdot s]) \\ &= \pi^{\sigma(i)}(b \mid \sigma \cdot s) = \pi^i(b|s).\end{aligned}$$

Thus the local policy is constant on every combined pointed orbit and defines one rule on $\mathcal{W}$. Conversely, a rule on $\mathcal{W}$ is unchanged under the elements (g, e) and (e, σ), so its induced policy is both G invariant and permutation invariant. This proves the policy correspondence. □

Proof of Theorem 3. Fix an agent i and recall its stabilizer H_i. Against jointly invariant opponents, agent i's best-response MDP is symmetric under $G \times H_i$. Indeed, reindexing the opponents' action sum by τ and applying first the permutation identities and then the state-only identities gives, for every $g \in G$, $\tau \in H_i$, and focal action b,

$$\begin{aligned}P^i_{\pi^{-i}}(L_g[\tau \cdot s'] \mid L_g[\tau \cdot s], b) &= P^i_{\pi^{-i}}(s' \mid s, b), \\ r^i_{\pi^{-i}}(L_g[\tau \cdot s], b) &= r^i_{\pi^{-i}}(s, b).\end{aligned}$$

The focal action remains b because G fixes all action labels and $\tau(i) = i$. Value iteration from zero therefore yields a $G \times H_i$-invariant optimal value and action-value function. Greedy selection with one fixed tie-breaking order gives a $G \times H_i$-invariant optimal stationary response.

The $G \times H_i$ orbits of states are exactly the fibers of w^i. One implication follows immediately from Eq. (9). For the converse, if $w^i(s) = w^i(s')$, then some (g, σ) maps (i, s) to (i, s'); hence $\sigma(i) = i$, so $\sigma \in H_i$ and $s' = L_g[\sigma \cdot s]$. The invariant optimal response is therefore of the form χ^i_w. Taking the maximum of these equalities over i proves Eq. (10). If μ is exchangeable, the relabeling argument in Lemma I.1 applies because a jointly invariant policy is permutation invariant; it shows that all agents have the same best-response gain. □

N Approximate Symmetry

This appendix states and proves the extension of Theorem 1 announced in Section 3.3: a measure of how far exact symmetry fails, and a bound showing that the same quotient, with orbit sums replaced by orbit averages, still tracks the unrestricted Nash gap up to an explicit error given by that measurement. Let G act on $\mathcal{S}$ by bijections as in Section 3.1, but assume nothing about P or $\vec{r}$, and put

$$\epsilon_P := \max_{g,s,a} \sum_{u \in \mathcal{S}} \left| P(L_g[u] \mid L_g[s], a) - P(u \mid s, a) \right|, \quad \epsilon_r := \max_{g,i,s,a} \left| r^i(L_g[s], a) - r^i(s, a) \right|. \tag{44}$$

These are definitions rather than hypotheses: every game has them, both are finite maxima computable from the model, and they vanish exactly when Definition 1 holds. The result below therefore needs no assumption on M at all, and is informative when Eq. (44) is small.

Representative independence of the definitions Eq. (3) now fails, so a choice must be made, and the right one is to average. Symmetrize M in place, keeping $\mathcal{S}$, $\mathcal{A}$, μ and γ fixed:

$$P_G(s' \mid s, a) := \frac{1}{|G|} \sum_{g \in G} P(L_g[s'] \mid L_g[s], a), \qquad r^i_G(s, a) := \frac{1}{|G|} \sum_{g \in G} r^i(L_g[s], a), \tag{45}$$

and write M_G for the resulting Markov game. Then M_G is exactly G-state-symmetric with rewards still in $[0, 1]$, and it lies within ϵ_P and ϵ_r of M in the norms of Eq. (44). Its quotient in the sense of Definition 2 need not be assembled from M_G, which would cost a row per state of $\mathcal{S}$ and defeat the reduction; instead it is the *orbit-averaged* quotient of M,

$$\overline{P}_G(\bar{s}' \mid \bar{s}, a) = \frac{1}{|\bar{s}|} \sum_{s \in \bar{s}} \sum_{u \in \bar{s}'} P(u \mid s, a), \qquad \overline{r}^i_G(\bar{s}, a) = \frac{1}{|\bar{s}|} \sum_{s \in \bar{s}} r^i(s, a), \tag{46}$$

with $\overline{\mu}$ the orbit sum of μ as before, so it is built directly from M with one row per orbit. When the rows of an orbit agree, Eq. (46) is Eq. (3).

Theorem N.1 (Approximate state quotienting)**.** *Let G act on $\mathcal{S}$ by bijections and let ϵ_P, ϵ_r be as in Eq.* (44)*. Then every $\bar{\pi} \in \Pi(\overline{M}_G)$ satisfies*

$$\left|\mathrm{NashGap}_M(\bar{\pi}^{\uparrow}) - \mathrm{NashGap}_{\overline{M}_G}(\bar{\pi})\right| \le \Delta_{\mathrm{sym}} := \frac{2\epsilon_r}{1-\gamma} + \frac{2\gamma\epsilon_P}{(1-\gamma)^2}. \tag{47}$$

The bound is two-sided and collapses to Theorem 1 at $\epsilon_P = \epsilon_r = 0$, so Eq. (47) relaxes that identity rather than replacing it. The proof keeps the exact theorem intact as a black box and adds two perturbation steps: the reduced gap equals $\mathrm{NashGap}_{M_G}(\bar{\pi}^{\uparrow})$ because M_G is exactly symmetric, and $\mathrm{NashGap}_{M_G}$ and $\mathrm{NashGap}_M$ differ by at most Δ_{sym} because every product policy's value moves by at most $\Delta_{\mathrm{sym}}/2$ between two games whose transitions and rewards are that close. Nothing is assumed of μ: since state symmetry constrains only P and r, the initial distribution is left alone and no initial-distribution term appears. The factors $(1-\gamma)^{-1}$ and $(1-\gamma)^{-2}$ are the usual simulation-lemma scalings, one horizon factor for a reward error and two for a transition error.

The same argument applies verbatim to agent permutations, since Theorem 2 is also used only as a black box; we omit the statement. Two things are lost relative to the exact case. Forming the orbit-averaged quotient costs $|G|$ terms per entry rather than one, and the MPG structure of Section 4.1 does not survive orbit averaging: since values are not linear in the kernel, M_G is an MPG only up to an additive $O(\Delta_{\mathrm{sym}})$ error, except for cooperative games, which stay cooperative MPGs. Any bound on $\mathrm{NashGap}_{\overline{M}_G}$, however obtained, still transfers to $\mathrm{NashGap}_M$ at the additive cost Δ_{sym}.

N.1 Proofs

Throughout the proofs, M and M' denote Markov games sharing $\mathcal{N}, \mathcal{S}, \mathcal{A}, \mu, \gamma$ and having rewards in $[0,1]$, and a prime marks a quantity computed in M': its transitions and rewards are P' and r'^i, its values V'^i_π.

Lemma N.1 (Symmetrization)**.** *Let G act on $\mathcal{S}$ by bijections, with ϵ_P, ϵ_r as in Eq.* (44)*. Then:*

(a) M_G of Eq. (45) *is a Markov game with $r^i_G \in [0,1]$ and is G-state-symmetric in the sense of Definition 1.*

(b) $\max_{s,a} \|P_G(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \le \epsilon_P$ and $\max_{i,s,a} \left|r^i_G(s,a) - r^i(s,a)\right| \le \epsilon_r$.

(c) The quotient of M_G in the sense of Definition 2 has the transitions and rewards Eq. (46)*, with $\overline{\mu}$ the orbit sum of μ.*

Proof. Part (a). For each g the map $s' \mapsto L_g[s']$ is a bijection of $\mathcal{S}$, so $\sum_{s'} P(L_g[s'] \mid L_g[s], a) = 1$ and hence $\sum_{s'} P_G(s' \mid s,a) = 1$; all terms are nonnegative, so $P_G(\cdot \mid s,a) \in \Delta(\mathcal{S})$. Each r^i_G is an average of values of r^i and so lies in $[0,1]$. For symmetry, fix $h \in G$ and use $L_g L_h = L_{gh}$ together with the reindexing $g \mapsto gh^{-1}$ of the group sum:

$$P_G(L_h[s'] \mid L_h[s], a) = \frac{1}{|G|} \sum_{g \in G} P(L_{gh}[s'] \mid L_{gh}[s], a) = P_G(s' \mid s, a),$$

and identically $r^i_G(L_h[s], a) = r^i_G(s,a)$. Since G fixes agents and actions, this is Definition 1.

Part (b). By the triangle inequality and Eq. (44),

$$\begin{aligned}\|P_G(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 &= \sum_{u \in \mathcal{S}} \left| \frac{1}{|G|} \sum_{g \in G} \left[P(L_g[u] \mid L_g[s], a) - P(u \mid s, a)\right] \right| \\ &\le \frac{1}{|G|} \sum_{g \in G} \sum_{u \in \mathcal{S}} |P(L_g[u] \mid L_g[s], a) - P(u \mid s, a)| \le \epsilon_P,\end{aligned}$$

and the same computation with r^i in place of P gives the reward bound.

Part (c). Fix $\bar{s}, \bar{s}'$ and $s \in \bar{s}$. Each L_g permutes $\bar{s}'$, so reindexing that orbit by $v = L_g[u]$ gives

$$\sum_{u \in \bar{s}'} P_G(u \mid s, a) = \frac{1}{|G|} \sum_{g \in G} \sum_{u \in \bar{s}'} P(L_g[u] \mid L_g[s], a) = \frac{1}{|G|} \sum_{g \in G} \sum_{v \in \bar{s}'} P(v \mid L_g[s], a).$$

By orbit–stabilizer the list $(L_g[s])_{g\in G}$ contains every element of $\bar{s}$ exactly $|G|\,/\,|\bar{s}|$ times, so the outer average equals the uniform average over $w \in \bar{s}$:

$$\sum_{u\in\bar{s}'} P_G(u \mid s,a) = \frac{1}{|\bar{s}|}\sum_{w\in\bar{s}}\sum_{v\in\bar{s}'} P(v \mid w,a).$$

The same argument applied to r_G^i gives the reward entry of Eq. (46). Both expressions are manifestly independent of the representative s, in accordance with Lemma A.1 applied to M_G, and $\overline{\mu}$ is unchanged because Definition 2 defines it from μ alone. □

Lemma N.2 (Value perturbation)**.** *Suppose* $\max_{s,a}\|P'(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \le \epsilon_P$ *and* $\max_{i,s,a}\left|r'^{\,i}(s,a) - r^i(s,a)\right| \le \epsilon_r$*. Then every stationary product policy* π *and every agent* i *satisfy*

$$\left\|V'^{\,i}_\pi - V^i_\pi\right\|_\infty \le \frac{\epsilon_r}{1-\gamma} + \frac{\gamma\epsilon_P}{(1-\gamma)^2} = \frac{\Delta_{\mathrm{sym}}}{2}. \tag{48}$$

Proof. Write $r^i_\pi(s) := \sum_a \pi(a \mid s) r^i(s,a)$ and $(P_\pi v)(s) := \sum_a \pi(a \mid s)\sum_{s'} P(s' \mid s,a)v(s')$, with $r'^{\,i}_\pi$ and P'_π defined from M' in the same way. The two value functions are the unique fixed points of $V^i_\pi = r^i_\pi + \gamma P_\pi V^i_\pi$ and $V'^{\,i}_\pi = r'^{\,i}_\pi + \gamma P'_\pi V'^{\,i}_\pi$. Writing $D := V'^{\,i}_\pi - V^i_\pi$ and subtracting,

$$D = \big(r'^{\,i}_\pi - r^i_\pi\big) + \gamma P'_\pi D + \gamma\big(P'_\pi - P_\pi\big)V^i_\pi.$$

Averaging over actions preserves the bounds $\|P'_\pi v\|_\infty \le \|v\|_\infty$ and, by Hölder's inequality applied statewise, $\|(P'_\pi - P_\pi)v\|_\infty \le \epsilon_P\|v\|_\infty$. Since $r^i \in [0,1]$ gives $\|V^i_\pi\|_\infty \le (1-\gamma)^{-1}$, taking supremum norms yields

$$\|D\|_\infty \le \epsilon_r + \gamma\|D\|_\infty + \frac{\gamma\epsilon_P}{1-\gamma},$$

and rearranging gives Eq. (48). □

Lemma N.3 (Gap perturbation)**.** *Suppose* $\left|V'^{\,i}_\pi(\mu) - V^i_\pi(\mu)\right| \le \eta$ *for every agent* i *and every product policy* π*. Then* $|\mathrm{NashGap}_{M'}(\pi) - \mathrm{NashGap}_M(\pi)| \le 2\eta$ *for every product policy* π*.*

Proof. Fix π. For each agent i and each local deviation $\widetilde{\pi}^i : \mathcal{S} \to \Delta(\mathcal{A}^i)$, both $(\widetilde{\pi}^i, \pi^{-i})$ and π are product policies, so

$$\left|\left[V'^{\,i}_{\widetilde{\pi}^i,\pi^{-i}}(\mu) - V'^{\,i}_\pi(\mu)\right] - \left[V^i_{\widetilde{\pi}^i,\pi^{-i}}(\mu) - V^i_\pi(\mu)\right]\right| \le 2\eta.$$

The two Nash gaps Eq. (1) are maxima of these two families over the same index set $(i,\widetilde{\pi}^i)$. A maximum of functions that differ pointwise by at most 2η differs by at most 2η, which is the claim. □

Proof of Theorem N.1. By Lemma N.1(a), M_G is G-state-symmetric, so Theorem 1 applies to it and gives, for every $\bar{\pi} \in \Pi(\overline{M}_G)$,

$$\mathrm{NashGap}_{M_G}(\bar{\pi}^\uparrow) = \mathrm{NashGap}_{\overline{M}_G}(\bar{\pi}).$$

The lift $\bar{\pi}^\uparrow$ is determined by the orbit map ρ alone, so it is the same policy in M and in M_G. By Lemma N.1(b) the two games satisfy the hypotheses of Lemma N.2 with the constants ϵ_P, ϵ_r of Eq. (44), so every product policy's value at μ moves by at most $\Delta_{\mathrm{sym}}/2$ between them. Lemma N.3 with $\eta = \Delta_{\mathrm{sym}}/2$ then gives

$$\left|\mathrm{NashGap}_M(\bar{\pi}^\uparrow) - \mathrm{NashGap}_{M_G}(\bar{\pi}^\uparrow)\right| \le \Delta_{\mathrm{sym}},$$

and combining the two displays proves Eq. (47). At $\epsilon_P = \epsilon_r = 0$ the definitions Eq. (45) give $M_G = M$ and $\Delta_{\mathrm{sym}} = 0$, recovering Theorem 1. □

O PROOFS OF THE SAMPLE-COMPLEXITY RESULTS

This section proves This section states and proves Theorem O.1, and proves Corollaries 2 and O.1. Within it, abbreviate $S := |\mathcal{S}|$ and $\bar{S} := |\overline{\mathcal{S}}|$. The first three lemmas are stated for a generic game so that all three results can share them. Throughout, X denotes a finite MPG $\langle \mathcal{N}, \mathcal{S}_X, \mathcal{A}_X, P_X, \vec{r}_X, \mu_X, \gamma \rangle$ with potential ϕ_X and total potential Φ_π^X. Let $\Phi_{\sup,X}$ be its finite bound from Eq. (36) in Appendix K. For a state quotient, part (a) of Lemma D.1 gives $\Phi_{\sup,\overline{M}} \le \Phi_{\sup,M}$.

The *potential MDP* of X has state space $\mathcal{S}_X$, action space $\mathcal{A}_X$, reward ϕ_X, kernel P_X, and discount γ; the value it assigns to a product policy π is the total potential Φ_π^X. Let

$$\Phi_\star^X(x) := \max_\pi \Phi_\pi^X(x)$$

denote its optimal value, the maximum ranging over all policies of that MDP, including history-dependent ones, and write $\Phi_\star^X(\mu_X) := \mathbb{E}_{x\sim\mu_X}[\Phi_\star^X(x)]$. A finite discounted MDP admits a stationary deterministic optimal policy, obtained by greedy selection with respect to its optimal action values; fix such a policy $\pi_\star$ for the potential MDP of X. For a transition kernel Q on $\mathcal{S}_X \times \mathcal{A}_X$, let Φ_π^Q denote the value functions of the potential MDP with Q in place of P_X, fix a stationary deterministic policy π_Q optimal for it, and set

$$\varepsilon_Q := \max_{(x,a)\in\mathcal{S}_X\times\mathcal{A}_X} \|Q(\cdot \mid x,a) - P_X(\cdot \mid x,a)\|_1, \qquad \eta_Q := \frac{\gamma\,\varepsilon_Q\,\Phi_{\sup,X}}{1-\gamma}.$$

Stationary deterministic policies of the potential MDP select one joint action per state and are therefore product policies of X.

Lemma O.1 (Simulation lemma for the potential MDP)**.** *Every product policy π of X satisfies*

$$\left\|\Phi_\pi^Q - \Phi_\pi^X\right\|_\infty \le \eta_Q.$$

Proof. For a stationary policy π of the potential MDP, define the averaged reward and the averaged kernel operators

$$\phi_{X,\pi}(x) := \sum_{a\in\mathcal{A}_X} \pi(a \mid x)\,\phi_X(x,a),$$
$$\big(P_{X,\pi}v\big)(x) := \sum_{a\in\mathcal{A}_X} \pi(a \mid x) \sum_{x'\in\mathcal{S}_X} P_X(x' \mid x,a)\,v(x'),$$

with Q_π defined from Q in the same way. The value functions are the unique fixed points of the Bellman equations

$$\Phi_\pi^X = \phi_{X,\pi} + \gamma\,P_{X,\pi}\Phi_\pi^X,$$
$$\Phi_\pi^Q = \phi_{X,\pi} + \gamma\,Q_\pi\Phi_\pi^Q.$$

Subtracting,

$$\Phi_\pi^Q - \Phi_\pi^X = \gamma\,Q_\pi\big(\Phi_\pi^Q - \Phi_\pi^X\big) + \gamma\big(Q_\pi - P_{X,\pi}\big)\Phi_\pi^X.$$

The averaged operator satisfies $\|Q_\pi v\|_\infty \le \|v\|_\infty$, and Hölder's inequality applied statewise after averaging over actions gives

$$\left\|\big(Q_\pi - P_{X,\pi}\big)v\right\|_\infty \le \varepsilon_Q\,\|v\|_\infty.$$

Taking supremum norms and rearranging,

$$\left\|\Phi_\pi^Q - \Phi_\pi^X\right\|_\infty \le \frac{\gamma\,\varepsilon_Q}{1-\gamma}\left\|\Phi_\pi^X\right\|_\infty.$$

By definition, $\left\|\Phi_\pi^X\right\|_\infty \le \Phi_{\sup,X}$. □

Lemma O.2 (Certainty-equivalent optimality transfer)**.** *For every $x \in \mathcal{S}_X$,*

$$\Phi_{\pi_Q}^X(x) \ge \Phi_\star^X(x) - 2\eta_Q.$$

Proof. The policies π_Q and $\pi_\star$ are stationary deterministic joint policies, hence product policies, so Lemma O.1 applies to both. For every x,

$$\begin{aligned}\Phi^X_{\pi_Q}(x) &\geq \Phi^Q_{\pi_Q}(x) - \eta_Q \\ &\geq \Phi^Q_{\pi_\star}(x) - \eta_Q \\ &\geq \Phi^X_{\pi_\star}(x) - 2\eta_Q = \Phi^X_\star(x) - 2\eta_Q.\end{aligned}$$

The middle inequality holds because π_Q is optimal for the Q-model potential MDP, whose optimal value dominates the Q-model value of every policy, in particular of $\pi_\star$. □

Lemma O.3 (Nash gap of an approximate potential maximizer)**.** *Let a product policy π of X satisfy $\Phi^X_\pi(x) \geq \Phi^X_\star(x) - \beta$ for every $x \in \mathcal{S}_X$ and some $\beta \geq 0$. Then* $\mathrm{NashGap}_X(\pi) \leq \beta$.

Proof. Fix an agent i and a stationary local deviation ν^i. The deviated policy (ν^i, π^{-i}) is a product policy and hence also a policy of the potential MDP of X, so its total potential is dominated by the optimal value:

$$\Phi^X_{\nu^i,\pi^{-i}}(x) \leq \Phi^X_\star(x) \qquad \text{for every } x \in \mathcal{S}_X.$$

Averaging over $x \sim \mu_X$ and applying the MPG identity of X at μ_X,

$$\begin{aligned}V^i_{\nu^i,\pi^{-i}}(\mu_X) - V^i_\pi(\mu_X) &= \Phi^X_{\nu^i,\pi^{-i}}(\mu_X) - \Phi^X_\pi(\mu_X) \\ &\leq \Phi^X_\star(\mu_X) - \big(\Phi^X_\star(\mu_X) - \beta\big) = \beta.\end{aligned}$$

Maximizing over ν^i and over agents bounds the Nash gap by β. □

Lemma O.4 (Uniform ℓ_1 concentration of the pooled estimator)**.** *Let M be a G-state-symmetric Markov game, and let $\widehat{\overline{P}}$ be the pooled estimator of Section 4.2: for each orbit–action pair $(\bar{s}, a)$, query the representative $(\iota(\bar{s}), a)$ n times, obtain next states $s'_1, \ldots, s'_n$, and set*

$$\widehat{\overline{P}}(\bar{s}' \mid \bar{s}, a) := \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\{s'_k \in \bar{s}'\}. \tag{49}$$

Fix $\varepsilon_P > 0$ and $\delta \in (0,1)$. If

$$n \geq \frac{2}{\varepsilon_P^2}\left(\big|\overline{\mathcal{S}}\big| \log 2 + \log \frac{\big|\overline{\mathcal{S}}\big|\,|\mathcal{A}|}{\delta}\right),$$

then, with probability at least $1-\delta$,

$$\max_{(\bar{s},a)\in\overline{\mathcal{S}}\times\mathcal{A}} \Big\|\widehat{\overline{P}}(\cdot \mid \bar{s}, a) - \overline{P}(\cdot \mid \bar{s}, a)\Big\|_1 \leq \varepsilon_P. \tag{50}$$

Proof. Fix a pair $(\bar{s}, a)$ and abbreviate $p := \overline{P}(\cdot \mid \bar{s}, a)$ and $\hat{p} := \widehat{\overline{P}}(\cdot \mid \bar{s}, a)$. For a draw $s' \sim P(\cdot \mid \iota(\bar{s}), a)$ and any subset $\mathcal{U} \subseteq \overline{\mathcal{S}}$,

$$\Pr\big[\rho(s') \in \mathcal{U}\big] = \sum_{\bar{s}'\in\mathcal{U}} \sum_{u\in\bar{s}'} P(u \mid \iota(\bar{s}), a) = p(\mathcal{U}), \tag{51}$$

so $\hat{p}(\mathcal{U})$ is the empirical mean of n independent Bernoulli variables with mean $p(\mathcal{U})$. For any two probability vectors on $\overline{\mathcal{S}}$,

$$\|\hat{p} - p\|_1 = 2\max_{\mathcal{U}\subseteq\overline{\mathcal{S}}} \big(\hat{p}(\mathcal{U}) - p(\mathcal{U})\big),$$

the maximum being attained at $\mathcal{U} = \{\bar{s}' : \hat{p}(\bar{s}') > p(\bar{s}')\}$. Hoeffding's inequality gives, for each fixed $\mathcal{U}$,

$$\Pr\left[\hat{p}(\mathcal{U}) - p(\mathcal{U}) \geq \frac{\varepsilon_P}{2}\right] \leq \exp\left(-\frac{n\varepsilon_P^2}{2}\right).$$

A union bound over the $2^{|\overline{\mathcal{S}}|}$ subsets, and then over the $\big|\overline{\mathcal{S}}\big|\,|\mathcal{A}|$ quotient pairs, bounds the failure probability of Eq. (50) by

$$\big|\overline{\mathcal{S}}\big|\,|\mathcal{A}|\; 2^{|\overline{\mathcal{S}}|} \exp\left(-\frac{n\varepsilon_P^2}{2}\right) \leq \delta$$

under the stated budget. □

One-line replacement for the appendix's transport construction — The pair-orbit space $\mathcal{Y}$, the representatives $\jmath(\bar{y})$, the permutations $\tau_{s,a}$, and the per-orbit budget of Corollary 2 are those of Section 4.2, stated there for M. The transported estimator queries each representative n times, forms its empirical row $\widehat{P}$, and sets

$$\widetilde{P}(s' \mid s, a) := \widehat{P}\big(\tau_{s,a}^{-1} \cdot s' \,\big|\, \jmath(y(s,a))\big)\,. \tag{52}$$

The proofs below treat a general finite agent-permutation-symmetric MPG X with state space $\mathcal{S}_X$, local action set $\mathcal{B}_X$, and the same objects subscripted by X ($\mathcal{Y}_X$, $\Phi_{\sup,X}$, and so on), so that the combined corollary at the end of this section can take $X = \overline{M}$; Corollary 2 is the case $X = M$. The following lemma is the concentration step.

Lemma O.5 (Uniform ℓ_1 concentration of the transported estimator)**.** *Let X be agent-permutation-symmetric, and let $\widetilde{P}$ be the transported estimator Eq.* (52) *with per-orbit budget n. Fix $\varepsilon_P > 0$ and $\delta \in (0,1)$. If*

$$n \geq \frac{2}{\varepsilon_P^2}\left(|\mathcal{S}_X| \log 2 + \log \frac{|\mathcal{Y}_X|}{\delta}\right),$$

then, with probability at least $1-\delta$,

$$\max_{(x,a)\in \mathcal{S}_X \times \mathcal{B}_X^N} \left\| \widetilde{P}(\cdot \mid x, a) - P_X(\cdot \mid x, a) \right\|_1 \leq \varepsilon_P. \tag{53}$$

Proof. Fix a pair (x,a), write $\bar{y} := y(x,a)$, $(x_0,a_0) := \jmath(\bar{y})$, and $\tau := \tau_{x,a}$, so that $\tau \cdot (x_0,a_0) = (x,a)$. The transition symmetry identity of the agent-permutation-symmetric class, applied with the permutation τ, states that $P_X(\tau \cdot u \mid \tau \cdot x_0, \tau \cdot a_0) = P_X(u \mid x_0, a_0)$ for every u. Substituting $u = \tau^{-1} \cdot x'$ gives the row identity

$$P_X(x' \mid x, a) = P_X\big(\tau^{-1} \cdot x' \,\big|\, x_0, a_0\big). \tag{54}$$

Combining Eq. (52) and Eq. (54), and reindexing the sum over x' by the bijection $x' \mapsto \tau^{-1} \cdot x'$,

$$\begin{aligned}\left\| \widetilde{P}(\cdot \mid x, a) - P_X(\cdot \mid x, a) \right\|_1 &= \sum_{x' \in \mathcal{S}_X} \left| \widehat{P}_X\big(\tau^{-1} \cdot x' \mid x_0, a_0\big) - P_X\big(\tau^{-1} \cdot x' \mid x_0, a_0\big) \right| \\ &= \left\| \widehat{P}_X(\cdot \mid x_0, a_0) - P_X(\cdot \mid x_0, a_0) \right\|_1 .\end{aligned}$$

Thus Eq. (53) holds as soon as every representative row is ε_P-accurate in ℓ_1. The subset argument in the proof of Lemma O.4 applies verbatim to each representative row, now with a union bound over the $2^{|\mathcal{S}_X|}$ subsets of $\mathcal{S}_X$ and the $|\mathcal{Y}_X|$ representative rows, bounding the failure probability by

$$|\mathcal{Y}_X|\; 2^{|\mathcal{S}_X|} \exp\left(-\frac{n\varepsilon_P^2}{2}\right) \leq \delta$$

under the stated budget. □

The sufficient budgets are compared through the proxy

$$B^{\mathrm{gen}}(\epsilon, \delta; S_0, A_0, \Phi_0) := \frac{8\gamma^2\Phi_0^2}{(1-\gamma)^2\epsilon^2}\, S_0 A_0 \left(S_0 \log 2 + \log \frac{S_0 A_0}{\delta}\right), \tag{55}$$

the sufficient number of queries to estimate every row of an S_0-state, A_0-action potential MDP to the accuracy used below. Pooling queries the $\bar{S}\,|\mathcal{A}|$ quotient pairs, with proxy $B^{\mathrm{gen}}(\epsilon, \delta; \bar{S}, |\mathcal{A}|, \Phi_{\sup})$; the same pipeline run directly on M (trivial group $G = \{e\}$) has proxy $B^{\mathrm{gen}}(\epsilon, \delta; S, |\mathcal{A}|, \Phi_{\sup})$, and

$$\frac{B^{\mathrm{gen}}(\epsilon, \delta; \bar{S}, |\mathcal{A}|, \Phi_{\sup})}{B^{\mathrm{gen}}(\epsilon, \delta; S, |\mathcal{A}|, \Phi_{\sup})} = \frac{\bar{S}}{S}\,\frac{\bar{S}\log 2 + \log(\bar{S}\,|\mathcal{A}|/\delta)}{S\log 2 + \log(S\,|\mathcal{A}|/\delta)} \leq \frac{\bar{S}}{S}, \tag{56}$$

the second factor being at most one since $S_0 \mapsto S_0 \log 2 + \log(S_0 A_0/\delta)$ is nondecreasing; when the linear terms dominate, the ratio is $(\bar{S}/S)^2$.

State-quotient theorem, demoted from the main text —

Theorem O.1 (Generative-model state-quotient guarantee)**.** *Let M be a G-state-symmetric MPG with $\gamma\Phi_{\sup} > 0$, and fix $\epsilon > 0$ and $\delta \in (0,1)$. If*

$$n \geq \frac{8\gamma^2\Phi_{\sup}^2}{(1-\gamma)^2\epsilon^2}\left(\left|\overline{\mathcal{S}}\right|\log 2 + \log\frac{\left|\overline{\mathcal{S}}\right||\mathcal{A}|}{\delta}\right), \tag{57}$$

then $\mathrm{NashGap}_M(\bar{\pi}_{\mathrm{ce}}^{\uparrow}) \leq \epsilon$ with probability at least $1-\delta$, where $\bar{\pi}_{\mathrm{ce}}$ is the certainty-equivalent policy for the potential MDP of $\overline{M}$ with kernel Eq. (49)*, using $n\left|\overline{\mathcal{S}}\right||\mathcal{A}|$ queries in total. Estimating every row of M needs the same budget with $|\mathcal{S}|$ in place of $\left|\overline{\mathcal{S}}\right|$, so pooling saves a factor $\left|\overline{\mathcal{S}}\right| / |\mathcal{S}| \leq |G|^{-1}$, and its square when the linear state terms dominate.*

Proof of Theorem O.1. Instantiate $X := \overline{M}$ and $Q := \widehat{\overline{P}}$, so that $\pi_Q = \bar{\pi}_{\mathrm{ce}}$. Since $\Phi_{\sup,\overline{M}} \leq \Phi_{\sup}$, it is valid to use the upper bound $\Phi_{\sup}$ below. Set

$$\varepsilon_P := \frac{(1-\gamma)\epsilon}{2\gamma\Phi_{\sup}}.$$

The budget Eq. (57) equals the budget of Lemma O.4 at this ε_P, so Eq. (50) holds with probability at least $1-\delta$. On that event, $\varepsilon_Q \leq \varepsilon_P$, hence $2\eta_Q \leq \epsilon$, and Lemmas O.2 and O.3 give

$$\mathrm{NashGap}_{\overline{M}}(\bar{\pi}_{\mathrm{ce}}) \leq 2\eta_Q \leq \epsilon.$$

By Eq. (4) in Theorem 1,

$$\mathrm{NashGap}_M\left(\bar{\pi}_{\mathrm{ce}}^{\uparrow}\right) = \mathrm{NashGap}_{\overline{M}}(\bar{\pi}_{\mathrm{ce}}) \leq \epsilon,$$

which proves the guarantee of Theorem O.1. Multiplying the per-pair budget Eq. (57) by the $\left|\overline{\mathcal{S}}\right||\mathcal{A}|$ queried pairs gives the proxy $B^{\mathrm{gen}}(\epsilon,\delta;\left|\overline{\mathcal{S}}\right|,|\mathcal{A}|,\Phi_{\sup})$ of Eq. (55) (with the ceiling per pair for the literal integer count); the no-symmetry proxy and the ratio Eq. (56) follow by running the same pipeline with the trivial group $G = \{e\}$, for which the quotient is M itself. □

We now prove Corollary 2 of Section 4.2, record the pooling ratio it implies, and derive the combined result.

Replacement for the appendix restatement of the permutation theorem —

Proposition O.1 (Pooling ratio for permutation symmetry)**.** *For a finite agent-permutation-symmetric MPG X as above, the total budget of Corollary 2 is, up to integer rounding,*

$$B^{\mathrm{pool}}(\epsilon,\delta;X) := \frac{8\gamma^2\Phi_{\sup,X}^2}{(1-\gamma)^2\epsilon^2}\,|\mathcal{Y}_X|\left(|\mathcal{S}_X|\log 2 + \log\frac{|\mathcal{Y}_X|}{\delta}\right). \tag{58}$$

Relative to estimating all $|\mathcal{S}_X|\,|\mathcal{B}_X|^N$ rows with the baseline proxy Eq. (55)*,*

$$\frac{B^{\mathrm{pool}}(\epsilon,\delta;X)}{B^{\mathrm{gen}}(\epsilon,\delta;|\mathcal{S}_X|,|\mathcal{B}_X|^N,\Phi_{\sup,X})} = \frac{|\mathcal{Y}_X|}{|\mathcal{S}_X|\,|\mathcal{B}_X|^N}\,\frac{|\mathcal{S}_X|\log 2 + \log(|\mathcal{Y}_X|/\delta)}{|\mathcal{S}_X|\log 2 + \log(|\mathcal{S}_X|\,|\mathcal{B}_X|^N/\delta)} \leq \frac{|\mathcal{Y}_X|}{|\mathcal{S}_X|\,|\mathcal{B}_X|^N}. \tag{59}$$

Proof of Corollary 2 and Proposition O.1. Take $Q := \widetilde{P}$, so that $\pi_Q = \pi_{\mathrm{ce}}$, and set

$$\varepsilon_P := \frac{(1-\gamma)\epsilon}{2\gamma\Phi_{\sup,X}}.$$

The state-quotient argument applies with Lemma O.5 in place of Lemma O.4 and the budget of Corollary 2 in place of Eq. (57). On the event Eq. (53), $\varepsilon_Q \leq \varepsilon_P$ and $2\eta_Q \leq \epsilon$, so Lemmas O.2 and O.3 give

$$\mathrm{NashGap}_X(\pi_{\mathrm{ce}}) \leq 2\eta_Q \leq \epsilon.$$

Multiplying the per-orbit budget of Corollary 2 by the $|\mathcal{Y}_X|$ queried orbits gives the continuous proxy Eq. (58); taking the ceiling per orbit gives the corresponding literal integer count.

For Proposition O.1, the untied baseline is the proxy Eq. (55) with state count $|\mathcal{S}_X|$ and action count $|\mathcal{B}_X|^N$, and with potential bound $\Phi_{\sup,X}$. The equality in Eq. (59) is direct substitution into Eq. (55) and Eq. (58). The inequality follows because the orbit map $(x,a) \mapsto y(x,a)$ is onto, so $|\mathcal{Y}_X| \leq |\mathcal{S}_X|\,|\mathcal{B}_X|^N$ and the second factor is at most one. □

Corollary O.1 (Combined generative-model pooling)**.** *Let M be an MPG with commuting symmetries (Definition 4), and let $\mathcal{Y}_{\overline{M}}$ be the state–joint-action orbit space of the agent-permutation-symmetric quotient. Query a representative in M, record the next-state orbit, and apply transported estimation on $\overline{M}$; let $\bar{\pi}_{\mathrm{ce}}$ denote the resulting certainty-equivalent policy. If $\gamma\Phi_{\mathrm{sup}} > 0$ and*

$$n \geq \frac{8\gamma^2\Phi_{\mathrm{sup}}^2}{(1-\gamma)^2\epsilon^2}\left(\left|\overline{\mathcal{S}}\right|\log 2 + \log\frac{|\mathcal{Y}_{\overline{M}}|}{\delta}\right),$$

then the lifted certainty-equivalent policy satisfies $\mathrm{NashGap}_M(\bar{\pi}_{\mathrm{ce}}^{\uparrow}) \leq \epsilon$ with probability at least $1-\delta$, using $n\,|\mathcal{Y}_{\overline{M}}|$ queries.

Proof of Corollary O.1. By Lemma M.1, $\overline{M}$ is agent-permutation-symmetric under the induced action $\sigma \cdot \bar{s} = \rho(\sigma \cdot s)$, and by Lemma D.1 it is an MPG with $\Phi_{\mathrm{sup},\overline{M}} \leq \Phi_{\mathrm{sup}}$. Its generative model is realized through M: querying a representative quotient pair means sampling $s' \sim P(\cdot \mid \iota(\bar{s}), a)$ and recording the orbit $\rho(s')$, which by Eq. (51) is an exact draw from $\overline{P}(\cdot \mid \bar{s}, a)$. Corollary 2 applied to $X = \overline{M}$, whose next-state space is $\overline{\mathcal{S}}$ and whose pair-orbit space is $\mathcal{Y}_{\overline{M}}$, then gives $\mathrm{NashGap}_{\overline{M}}(\bar{\pi}_{\mathrm{ce}}) \leq \epsilon$ with probability at least $1-\delta$ under the stated per-orbit budget, with total budget $n\,|\mathcal{Y}_{\overline{M}}|$. Finally, Eq. (4) transfers the bound: $\mathrm{NashGap}_M(\bar{\pi}_{\mathrm{ce}}^{\uparrow}) = \mathrm{NashGap}_{\overline{M}}(\bar{\pi}_{\mathrm{ce}}) \leq \epsilon$. □

P Experiment Hyperparameters

All runs use exact tabular gradients from the uniform policy ($\theta = 0$ for softmax, the uniform distribution for direct), evaluate the true Nash gap every 20 iterations, and compare the parameterizations at a common step size. Table 1 lists the settings; the generative-model results are means over 10 sampling seeds, and the optimization curves are single deterministic runs.

Table 1: Experiment hyperparameters.

Hyperparameter	**Value**
Discount γ	0.9
Ring MPG	$N = 3$ agents, cycle C_8, move-success probability 0.9
Cyclic-pursuit	2 players, cycle C_{10}, deterministic moves
Softmax PG (ring)	log-barrier, step $\eta = 5$, $\lambda = 10^{-4}$, 4000 iterations
Direct PG (ring)	projected, step $\eta = 0.5$, 2500 iterations
OGDA softmax (zero-sum)	step $\eta = 0.3$, 6000 iterations
OGDA direct (zero-sum)	projected, step $\eta = 0.02$, 6000 iterations
Generative samples per orbit n	$\{2, 4, 8, 16, 32, 64, 128\}$
Generative seeds	10 (mean $\pm$ standard deviation)