Playing Markov Games Without Observing Payoffs

Daniel Ablin
Department of Electrical Engineering
Tel Aviv University
Tel Aviv, Israel
danielablin@mail.tau.ac.il

Alon Cohen
Department of Electrical Engineering
Tel Aviv University
Tel Aviv, Israel
alonco@tauex.tau.ac.il

Abstract

Optimization under uncertainty is a fundamental problem in learning and decision-making, particularly in multi-agent systems. Previously, Feldman, Kalai, and Tennenholtz [2010] demonstrated the ability to efficiently compete in repeated symmetric two-player matrix games without observing payoffs, as long as the opponent's actions are observed. Extending this capability to the Markovian setting remains an open problem. In this paper, we introduce and formalize a new class of zero-sum symmetric Markov games, which extends the notion of symmetry from matrix games to the Markovian setting. We prove that a learner observing only the opponent's action sequence, without access to payoff information, can successfully compete against an adversary possessing complete knowledge of the game. We formalize three distinct notions of symmetry in this domain and reveal a surprising structural hierarchy: the most holistic definitions of symmetry impose restrictive constraints that actually simplify the learning landscape, whereas the "simplest" definition represents the most general and challenging setting. We provide polynomial-time algorithms for all three settings that achieve sublinear regret. Crucially, we demonstrate that despite the complex Markovian dynamics, a simple strategy of locally mimicking the opponent's actions suffices to guarantee robustness. This finding significantly broadens the class of games where robust learning is possible under severe informational disadvantage, proving that knowledge of the transition laws is not required to force a draw.

1 Introduction

In unfamiliar or competitive settings, imitation often emerges as a natural coping tactic. In rapidly evolving online markets, where advertising tactics can shift daily, newer companies frequently align their campaigns with those of more seasoned competitors. Assessing the success of a marketing effort is challenging, so inexperienced firms may gain by replicating the strategies of those with greater market knowledge. Lacking the time or budget for extensive research, these newcomers must act decisively to stay competitive against entrenched players. However, imitation carries significant risk: executed well, it can lead to outsize gains with minimal investment; executed poorly, it can result in outcomes worse than random chance. This problem is even greater if one considers the scenario where a company uses a long-range advertising campaign where each advertising strongly dependent on previous tactics. To study this behavior, we frame it as a strategic interaction where the objectives of the competitor may conflict with your own. Specifically, we consider a zero-sum symmetric game.

Previously, Feldman et al. [2010] studied repeated play of finite zero-sum symmetric matrix games. They considered a scenario where the learner has no information about the game and does not observe any payoff. The learner is only aware of the available actions and observes the actions played by the adversary. On the other hand, the adversary may have complete prior knowledge of the payoff matrix. They demonstrated a simple *copycat strategy* that guarantees, against any opponent's strategy:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

$\mathbb{E}[|\sum_{t=1}^{T} u(a_t^1, a_t^2)|] = O(n\sqrt{T})$, where T is the number of rounds, u is the payoff function, a_t^1, a_t^2 are the actions of the learner and adversary on round t, respectively, and n is the number of actions.

In our work we ask the following natural question: can the *copycat strategy* of Feldman et al. [2010] be extended to Markov games, where the current state depends on the history of actions? In such a setting, how can a learner effectively mimic an adversary and ensure sublinear return? Crucially, we operate under a severe informational disadvantage. We assume the learner must navigate the complex transition dynamics without possessing prior knowledge of the game's parameters, neither the payoff function nor the underlying transition laws are known. The learner must rely solely on observing the actions played by the adversary.

To address this, our first challenge is to formalize the concept of a zero-sum symmetric Markov game. This requires capturing a core intuition:

Fundamentally, symmetry implies that swapping roles and strategies perfectly swaps the players' payoffs, which, in a zero-sum game, simply negates the expected return.

Following this intuition, we explore three distinct notions of symmetry in Markov games.

When transitioning from static matrix games to Markov games, the most natural formulation of symmetry treats the game's overall expected value as the swapped payoff. This leads to our first definition, *symmetry with respect to Markov policies* (MSG), where we require that the game's overall value remains symmetric over the entire horizon. Because this holistic approach does not enforce constraints on any specific local state, it is seemingly the broadest and least restrictive class of games.

Next, once value symmetry is established for Markov policies, an immediate follow-up question arises: what happens if the symmetry is for non-Markovian strategies? This brings us to our second definition, *symmetry with respect to history-dependent policies* (HSG).

Finally, we consider a much more restrictive environment: enforcing symmetry locally at every single step. This leads to our third definition, *per-state symmetric games* (SSG), where every individual state corresponds to a symmetric zero-sum matrix game. We note that to the best of our knowledge, even the restrictive SSG is the most general formulation of perfect symmetry in Markov games, as previous literature typically imposes additional restrictions on the transition dynamics.

At first glance, this progression-from holistic value symmetry (MSG) to strict local symmetry (SSG)-seems to move from broad, complex game classes to a highly restrictive case. Nevertheless, as we discover and structurally prove in this work, this intuitive hierarchy is completely backwards: MSGs and HSGs are special cases of SSGs! Our result follows from carefully unfolding the global symmetry in MSGs and HSGs into a local one.

In light of this surprising result, we aim to design an efficient learning algorithm that works across all three settings. Unlike symmetric matrix games, where a learner can simply mimic the opponent's actions, Markov games introduce complex transition dynamics that the learner must navigate. We nevertheless provide polynomial-time algorithms for these settings while overcoming the aforementioned challenges. To the best of our knowledge, we are the first to demonstrate tractable learning in Markov games without observing payoffs, restrictive opponents, or requiring auxiliary information.

1.1 Related work

Blackwell Approachability. Blackwell Approachability is an extension of Von Neumann's minimax theorem [von Neumann, 1928] to vector-valued payoffs. The objective is to ensure, in repeated play, that the average payoff vector converges to a desirable convex target set K. Unlike in the scalar case, fixed (oblivious) strategies may fail to guarantee convergence. Blackwell's celebrated *Approachability Theorem* [Blackwell, 1956] shows that in the repeated game setting, an adaptive strategy that reacts to the opponent's play can still guarantee convergence to K.

Prior work by Abernethy et al. [2011], Shimkin [2016] shows that Blackwell Approachability and Online Convex Optimization are algorithmically equivalent in the context of repeated games. Specifically, any algorithm that achieves approachability can be transformed into an online algorithm with low regret, and vice versa.

Symmetric Markov Games. Several recent papers have studied how to formalize symmetry in Markov decision processes (MDPs) and Markov games. Zinkevich and Balch [2001] introduce

a notion of symmetry in MDPs and show how it can be exploited for more effective learning in single-agent as well as multi-agent systems. In their formulation, two states are symmetric if they have symmetric actions, and two actions are symmetric if they lead to symmetric outcomes. They further extend this idea to symmetric MDPs, requiring both the payoff and transition functions to be symmetric. Zinkevich [2006] broadens this perspective to settings involving multiple agents.

Another line of work formulates symmetry at the level of payoffs in Markov games. Flesch et al. [2013] analyze a two-player Markov game in which the payoff functions are symmetric: if the players' actions at a given state are exchanged, then their payoffs are swapped as well. A similar approach is taken by Yongacoglu et al. [2023], who consider a multi-player setting and require invariance under arbitrary permutations of the action sets. Both frameworks are flexible enough to capture zero-sum games. A key limitation, however, is that they assume transition dynamics are insensitive to the ordering of players' actions, i.e., swapping the players does not alter the state dynamics, which restricts the generality of the resulting games.

Learning in Markov games. Markov games pose significant theoretical and algorithmic challenges, particularly in adversarial and partially observable settings. Prior work identified strong hardness results for core learning objectives in this domain. For instance, Bai et al. [2020] prove that computing a best response policy in zero-sum Markov games when facing an adversarial opponent is at least as difficult as the well-known parity with noise problem, which is widely believed to be computationally intractable.

A closely related special case that has been extensively studied is that of MDPs with known transitions but adversarial or unknown rewards. Such an MDP can be interpreted as a Markov game in which the adversary acts as a second player whose "actions" either modify the state payoffs or, equivalently, redirect transitions to payoff-altered versions of the next state. This line of research (e.g., Even-Dar et al., 2009, Zimin and Neu, 2013, Jin and Luo, 2020) primarily focuses on scenarios where the payoff function may vary over time. In a related line of work, Tian et al. [2021], Liu et al. [2022] establish tight lower bounds on regret in the online learning setting, showing that no algorithm can achieve better than $\Omega(\min\{\sqrt{2^H T}, T\})$ regret in episodic zero-sum Markov games with unknown transitions, even when the adversary is limited to using Markov strategies.

Despite these limitations, recent advances have uncovered tractable settings under more cooperative assumptions. For example, Erez et al. [2023], Cai et al. [2024], Cui et al. [2023], Yang and Ma [2022], Jin et al. [2021] design decentralized algorithms for Markov games that achieve sublinear regret, assuming all players employ the same proposed learning rule. Their approach requires no communication between agents and guarantees convergence to a (coarse) correlated equilibrium.

2 Preliminaries

2.1 Markov games

A two-player zero-sum finite horizon Markov game is defined by the quintuple (S, A, H, P, u), where $H \geq 1$ is the time horizon of the game. S is a finite set of states partitioned as $S = \bigsqcup_{h=1}^{H} S_h$, where $S_1 = \{s_1\}$. Without loss of generality, we assume that all states in S are reachable from the initial state s_1 under some pair of policies, as unreachable states do not affect the game's value or dynamics. $A = \{1, \ldots, n\}$ is a finite set of actions available to both players; $P(\cdot \mid s_h, a^1, a^2) \in \Delta(S_{h+1})$ is the probability distribution over the next state when the players 1 and 2 take actions a^1 and a^2 respectively at state $s_h \in S_h$. Finally, $u : S \times A \times A \to [-1, 1]$ is the payoff function for player 1. As the game is zero-sum, $-u$ is the payoff function for player 2.

At every time step $h = 1, \ldots, H$, the players observe the current state s_h, then choose actions a_h^1, a_h^2 respectively. The players choose their actions independently at random according to their policies. Then player 1 gains $u(s_h, a_h^1, a_h^2)$ and player 2 gains $-u(s_h, a_h^1, a_h^2)$, and the game transitions to the next state s_{h+1} drawn from $P(\cdot \mid s_h, a_h^1, a_h^2)$. The game ends after time step H.

A history $z_h = (s_1, a_1^1, a_1^2, \ldots, s_{h-1}, a_{h-1}^1, a_{h-1}^2, s_h)$ is the sequence of states and actions generated by the players up to time h. We denote by Z_h the set of histories up to time h. A (history-dependent) policy for player i is a function $\pi^i : Z_h \mapsto \Delta(A)$ that maps each history z_h to a probability distribution from which player i draws their action when observing z_h. A policy is Markov if it is a function of only the current state, namely $\pi^i : S \mapsto \Delta(A)$.

We also introduce the concept of the *value function*, which captures the expected cumulative payoff from a given state onward and will be fundamental in defining optimality and proving performance guarantees. For policies π^1, π^2 of the two players, time step h and state $s \in S_h$, the *value function* is defined as: $V^{\pi^1,\pi^2}(s) \coloneqq \mathbb{E}^{\pi^1,\pi^2}\left[\sum_{h'=h}^{H} u(s_{h'}, a^1_{h'}, a^2_{h'}) \,\middle|\, s_h = s\right]$, where the expectation is taken over the randomness in the actions taken by the policies and the transition dynamics. For Markov policies, the value function satisfies the Bellman equations (see, e.g., Puterman, 1990):

$$V^{\pi^1,\pi^2}(s) = \mathbb{E}^{\pi^1,\pi^2}\left[u(s, a^1_h, a^2_h) + \sum_{s' \in S_{h+1}} P(s' \mid s, a^1_h, a^2_h) V^{\pi^1,\pi^2}(s') \,\middle|\, s_h = s\right].$$

We define the value of the game as $V^\star = \max_{\pi^1} \min_{\pi^2} V^{\pi^1,\pi^2}(s_1)$. The value is guaranteed by a Markov safety level policy $\pi^\star$ for player 1, that is $V^{\pi^\star,\pi^2}(s_1) \geq V^\star$ for any strategy π^2 for player 2. Moreover, there is a polynomial-time dynamic programming algorithm that returns such $\pi^\star$ [Wal and Wessels, 1976].

2.2 Online linear optimization

Online linear optimization is a sequential game between a learner and the environment played for T rounds. Before the game begins, the environment privately generates a sequence of linear loss functions $l_1, \ldots, l_T : U \mapsto \mathbb{R}$ where $U \subseteq \mathbb{R}^m$ is a compact convex set. At every round t, the learner predicts $u_t \in U$. Only then, the environment reveals l_t and the learner suffers $l_t(u_t)$. The regret of the learner is defined as: $R(T) \coloneqq \sum_{t=1}^{T} l_t(u_t) - \min_{u \in U} \sum_{t=1}^{T} l_t(u)$. A learning algorithm is said to be consistent if its regret grows sub-linearly with time, i.e., $R(T) = o(T)$. This ensures that, on average, the algorithm performs asymptotically as well as the best fixed strategy in hindsight.

Online Gradient Descent (OGD). OGD refers to the online optimization algorithm that starts from some $u_1 \in U$, then on every round t updates: (1) $u'_{t+1} = u_t - \eta\, l_t$ where $\eta > 0$ is the learning rate; (2) sets u_{t+1} to be the Euclidean projection of u'_{t+1} onto U. For a specific choice of η, OGD achieves regret bounded by $O(GD\sqrt{T})$ when $\|l_t\| \leq G$ for all $t = 1, \ldots, T$ and $\max_{u,u' \in U} \|u - u'\| \leq D$ (see, e.g., Hazan, 2022). Note that OGD operates in time complexity of $O(T(m + Q))$, where Q is the time to compute the projection on U. Thus if Q is polynomial in the problem parameters, the OGD runs in polynomial time as well.

3 Problem definition and main results

We consider the problem of learning a two-player zero-sum Markov symmetric game by repeated play. We operate under a severe informational disadvantage: the learner never observes the game's payoff, only the adversary's actions, and assumes no prior knowledge of the transition probability function P. In contrast, we assume the adversary can be omniscient, possessing complete prior knowledge of both the transition dynamics P and the payoff function u.

The game is played for T episodes. At the beginning of each episode t, both players choose policies π^1_t, π^2_t, that are then used to play the game in the current episode. While the opponent can choose any, possibly history-dependent, policy, we restrict the learner to select only Markov policies to facilitate efficient learning. Thereafter, both players observe the trajectory sampled from the game's dynamics and the actions of their opponent. Crucially, only the adversary can calculate the incurred payoff. The goal of the learner is to guarantee that their own payoff approaches the value of the game, which is 0.

Formally, the learner minimizes $\mathbb{E}[|\sum_{t=1}^{T} V^{\pi^1_t,\pi^2_t}(s_1)|]$. We aim for a bound that is sublinear in T, implying the average payoff converges to zero. The concept of a zero-sum symmetric Markov game, at its core, must capture the intuition that swapping roles and strategies perfectly swaps the players' payoffs, which, in a zero-sum game, simply negates the expected return. Following this intuition, we consider three types of symmetry in Markov games. We begin by formalizing the most natural and holistic extension of symmetry to the Markovian setting. By treating the game in its entirety as symmetric, we require that the expected cumulative return perfectly reflects role-swapping over the space of Markov policies.

Assumption 1 (Symmetry w.r.t. Markov policies; MSG)**.** *A zero-sum Markov game is MSG if for every pair of Markov policies π^1, π^2 it holds that $V^{\pi^1,\pi^2}(s_1) = -V^{\pi^2,\pi^1}(s_1)$.*

Next, we consider the natural generalization of the MSG definition, where symmetry must hold even for history-dependent policies.

Assumption 2 (Symmetry w.r.t. history-dependent policies; HSG)**.** *A zero-sum Markov game is HSG if for every two policies π^1, π^2, possibly history-dependent, it holds that $V^{\pi^1,\pi^2}(s_1) = -V^{\pi^2,\pi^1}(s_1)$.*

Finally, while MSG and HSG enforce symmetry on the overall value of the game, one might ask: what if we require each individual state to be symmetric?

Assumption 3 (Per-state symmetric game; SSG)**.** *A zero-sum Markov game is SSG if $u(s, a^1, a^2) = -u(s, a^2, a^1)$ for all $s \in S$ and $a^1, a^2 \in A$. Namely, the payoff matrix in each state corresponds to a zero-sum symmetric matrix game.*

3.1 Our results

Our analysis reveals a counter-intuitive hierarchy across the three symmetry classes. One might naturally assume that "holistic" notions of symmetry (MSG and HSG), which enforce symmetry over policies rather than local states, would yield a richer, more general class of games. However, we establish the exact opposite: holistic symmetry requirements impose such severe constraints on the game structure that they effectively simplify the learning landscape. We detail this hierarchy below.

First, for **MSG** environments (Assumption 1), defining symmetry holistically over full policies introduces complex temporal dependencies that mathematically couple the entire state space. This drastically increases the complexity of the game, rendering direct optimization or equilibrium analysis computationally prohibitive, if not entirely intractable.

To overcome this, we develop a structural reduction that disentangles the game's complex transition dynamics. We show that by identifying the correct transformation, we can recursively "peel back" the horizon, revealing that every layer of an MSG effectively decomposes into independent skew-symmetric matrix games. We demonstrate that this transformation preserves the game's value for any pair of strategies. Formally, we prove the following:

Theorem 1. *Any MSG can be efficiently transformed into an SSG by recursively applying the Expected Payoff Redistribution (EPR) reduction, without changing the expected cumulative payoff for any pair of policies.*[1]

This result (detailed in Section 4) implies that, in practice, the learner can simply act as if they are playing against the modified SSG. By doing so, they seamlessly preserve all subsequent baseline learning guarantees while entirely bypassing the complex, temporally coupled dynamics of the original MSG.

Next, for **HSG** environments (Assumption 2), we demonstrate that the symmetry constraint imposes a strong structural restriction on the game. In particular, while the game may continue through arbitrary stochastic transitions as long as both players choose identical actions, once the players choose different actions, their subsequent choices cannot affect the expected remaining payoff. Formally, we prove the following:

Theorem 2. *Any HSG can be transformed into a value-equivalent SSG that terminates upon the first disagreement between the players.*

This result (Section 6) reveals that HSGs are structurally equivalent to games that terminate upon the first disagreement between the players. Importantly, the game may still involve complex stochastic transitions and multiple states before this disagreement. Consequently, the transition-free learning algorithm developed for SSGs also applies to HSGs.

Finally, for **SSG** environments (Assumption 3), we establish the following baseline guarantee.

Theorem 3. *For any SSG, we present a learning algorithm that guarantees $\mathbb{E}[|\sum_{t=1}^{T} V^{\pi_t^1,\pi_t^2}(s_1)|] = O(n\sqrt{T|S|H})$ in $\mathrm{poly}(|S|, n, T)$ run-time.*

[1] Note that the EPR transformation accumulates payoffs over time steps, which may inflate them by up to a factor of H.

This result (detailed in Section 5) relies on a reduction to online linear optimization via Blackwell approachability. Moreover, our algorithm is implemented efficiently using the observation that local state symmetry allows the learner to decouple the Markov game.

To complement these upper bounds, we provide a matching lower bound for SSG games in Section B.

Proposition 1. *There exists an adversary and a payoff function for which* $\mathbb{E}[|\sum_{t=1}^{T} V^{\pi_t^1,\pi_t^2}(s_1)|] = \Omega(n\sqrt{T|S|H})$ *against any, possibly randomized, learner. This result confirms that our algorithms operate in the optimal regime with respect to the parameters of the game.*

Finally, to complete our theoretical analysis, we demonstrate the robustness of our framework by extending our results to general-sum games (Section 5.1) and environments with ϵ-approximate symmetry (Section G).

Significance of results. To the best of our knowledge, this work is the first to define symmetry holistically over the entire Markov game rather than restricting it to local states. We adopt this holistic approach because it faithfully captures the true underlying intuition of a zero-sum symmetric game: that swapping roles and strategies should perfectly negate the expected return.

At first glance, MSG may appear to define a broader and more challenging class than SSG, introducing local state asymmetries and mathematically coupled global constraints that render direct optimization computationally prohibitive, let alone when extended to HSG. This work reveals a counter-intuitive reality: our structural reductions (Theorems 1 and 2) prove that attempting to induce complexity through these holistic definitions neutralizes the impact of the transition dynamics. By collapsing these temporally coupled games into independent matrix games, we bypass highly complex algorithmic constructions. Consequently, our findings suggest that SSG is not merely the simplest definition, but structurally the most general.

This structural collapse profoundly expands the boundaries of symmetric MARL research. In the broad literature on finding equilibria in symmetric zero-sum games, the standard requirement for establishing a Markov Perfect Equilibrium (MPE) is strict, local, per-state symmetry for both payoffs and transitions (which constitutes a subset of our SSG). Nevertheless, our work reveals that SSG is arguably the most general definition of symmetry utilized in Markov game literature, and our results suggest that any prior work relying on stricter constraints might be able to drastically relax their assumptions. By porting their results through our transformation, the entire field's definition of symmetry can be potentially broadened to encompass holistic policy symmetry (e.g., MSG, HSG) while preserving fundamental equilibrium or regret guarantees. In summary, our structural collapse is not just a stepping stone for regret bounds, but also provides a powerful blueprint for simplifying complex sequential decision-making problems.

4 Copycat strategy in MSG

In this section, we prove Theorem 1. Our approach to learning under the MSG assumption (Assumption 1) focuses on exploiting the structural constraints induced by the symmetry of the value function.

The key distinction from SSG is that symmetry is now enforced on the expected cumulative payoff rather than locally at each state. The set of valid MSG payoff functions is defined as:

$$U = \left\{ u \in \mathbb{R}^{|S|\times n\times n} \ : \ |u(\cdot)| \leq 1, \text{ and } V^{\pi^1,\pi^2}(s_1) = -V^{\pi^2,\pi^1}(s_1), \forall \pi^1, \pi^2 \in \Pi_M \right\},$$

where Π_M denotes the set of all Markov policies.

The central challenge in the MSG setting is that while the game is globally zero-sum symmetric, individual states need not be skew-symmetric. This breaks the local guarantee of the Minimax strategy used in the matrix case. Intuitively, one might expect the MSG assumption to be a broad generalization of SSG. However, we demonstrate that this intuition is misleading: the holistic nature of MSG imposes constraints that some SSG games fail to satisfy. We illustrate this distinction with the counter-example in Section H.1.

We leverage this insight by analyzing the game from the latest timestep, by fixing the strategy up to the final state. This motivates our recursive reduction technique.

4.1 Expected Payoff Redistribution

We introduce the **Expected Payoff Redistribution (EPR)** transformation, which propagates expected future rewards into the current state to reveal the underlying symmetry of the future states.

Definition 1 (Expected Payoff Redistribution; EPR)**.** *Given a payoff function u with horizon $H \geq 2$, the EPR transformation on round h yields a new payoff $\hat{u}$ defined as:*

- $\hat{u}(s, a^1, a^2) = u(s, a^1, a^2) + \sum_{s' \in S_{h+1}} \mathbb{P}(s' \mid s, a^1, a^2) u(s', 1, 1)$ *for all* $s \in S_h$.
- $\hat{u}(s, \cdot, \cdot) = u(s, \cdot, \cdot) - u(s, 1, 1)$ *for all* $s \in S_{h+1}$.
- $\hat{u}(s, \cdot, \cdot) = u(s, \cdot, \cdot)$ *otherwise.*

Lemma 1 (EPR on MSG)**.** *Consider an MSG game with $H \geq 2$. Applying the EPR transformation (Definition 1) on round $H - 1$ yields $\hat{u}$ with the following properties:*

1. *At round H, the payoff matrix for every state is skew-symmetric.*
2. *If we truncate the last round, the resulting game of length $H - 1$ satisfies the MSG assumption (1).*

Proof Sketch. The full proof (deferred to Section D) relies on mathematically "calibrating" the game using a reference action pair. Because the MSG assumption mandates that $V^{\pi^1, \pi^2}(s_1) = -V^{\pi^2, \pi^1}(s_1)$ for *all* Markov policies, we can isolate the final step's payoff structure by evaluating specific, paired strategies. Specifically, we construct a baseline strategy where both players deterministically play a fixed action (e.g., action 1) upon reaching any state $s \in S_H$, while playing uniformly at random in all preceding states. By substituting these policies into the MSG condition and measuring the value of arbitrary actions at step H relative to this baseline, we effectively subtract out the influence of the complex transition history.

This algebraic isolation reveals a rigid structure in the final layer: for all actions, $u(s_H, a^1, a^2) + u(s_H, a^2, a^1) = 2u(s_H, 1, 1)$. While this final payoff matrix is not inherently skew-symmetric, it becomes strictly skew-symmetric once we subtract the baseline value of playing $(1, 1)$. The Expected Payoff Redistribution (EPR) transformation mechanically executes this by explicitly subtracting the $(1, 1)$ baseline from step H and redistributing it backward to step $H - 1$. This completely removes the "bias" at step H, forcing its local payoff matrix into a skew-symmetric form, while successfully propagating the expected future value backward without violating the overall MSG assumption for the truncated $H - 1$ game. □

Lemma 2. *For any payoff function u with $H \geq 2$, applying EPR (Definition 1) at any round $h < H$ preserves the value of the game. Specifically, for any strategies (possibly history-dependent) π^1, π^2: $V^{\pi^1, \pi^2}(s_1) = \hat{V}^{\pi^1, \pi^2}(s_1)$, where $\hat{V}$ is the value function under $\hat{u}$.*

Proof. Since $\hat{u}$ differs from u only at steps h and $h + 1$, it suffices to show that value conservation holds locally. For any $z \in Z_h$ ending at state s, let $z' = z \oplus (a^1, a^2, s')$ for each $s' \in S_{h+1}$:

$$
\begin{aligned}
V^{\pi^1, \pi^2}(z) &= \mathbb{E}^{\pi^1, \pi^2}\left[u(s, a^1, a^2) + \sum_{s' \in S_{h+1}} \mathbb{P}(s' \mid s, a^1, a^2) V^{\pi^1, \pi^2}(z') \right] \\
&= \mathbb{E}^{\pi^1, \pi^2}\left[u(s, a^1, a^2) + \sum_{s' \in S_{h+1}} \mathbb{P}(s' \mid s, a^1, a^2) (u(s', 1, 1) + \hat{V}^{\pi^1, \pi^2}(z')) \right] \\
&= \mathbb{E}^{\pi^1, \pi^2}\left[\hat{u}(s, a^1, a^2) + \sum_{s' \in S_{h+1}} \mathbb{P}(s' \mid s, a^1, a^2) \hat{V}^{\pi^1, \pi^2}(z') \right] = \hat{V}^{\pi^1, \pi^2}(z). \qquad \square
\end{aligned}
$$

4.2 Reducing MSG to SSG.

Lemma 3. *For any MSG game with payoff u, there exists an SSG game with payoff $\hat{u}$ such that for all policies (possibly history-dependent): $V^{\pi^1, \pi^2}(s_1) = \hat{V}^{\pi^1, \pi^2}(s_1)$.*

Proof. We prove this by recursive reduction. We start by applying the EPR transformation (Definition 1). Lemma 1 guarantees that the transformed game, when viewed as a game of horizon $H - 1$, retains the MSG property. Simultaneously, it ensures the payoff at the discarded H-th time step becomes strictly skew-symmetric.

We repeat this process $H-1$ times. At each step, we **decouple** the final time step, transforming its local payoff matrices into skew-symmetric form, and **propagate** the residual asymmetry to the preceding layer. Lemma 2 ensures that the value function remains invariant throughout. In the final step, the remaining payoff matrix at s_1 must itself be skew-symmetric to satisfy the MSG assumption (1). The result is a fully skew-symmetric game (SSG) that is equivalent to the original MSG. □

Consequently, from the learner's perspective, the MSG class is structurally equivalent to SSG. The learner can apply the SSG algorithm directly to any MSG instance without modification. Building on this foundational reduction, the next section formally analyzes the SSG setting (Assumption 3).

5 Copycat strategy in SSG

In this section, we prove Theorem 3 and present our strategy for learning in Per-state Symmetric Games (SSG) under Assumption 3. In this setting, every state $s \in S$ encapsulates a symmetric zero-sum matrix game. While Section F details a specialized alternative approach that runs the matrix copycat algorithm simultaneously at every state, we focus here on a broader, general framework.

Notably, in Section C we naturally frame the copycat problem in matrix games [Feldman et al., 2010] as a Blackwell approachability problem, efficiently solvable via online optimization. Here, we generalize this scheme to Markov games. Our formal analysis relies on direct reductions to online optimization rather than explicitly invoking Blackwell approachability, which drastically simplifies our presentation. We now formally analyze the SSG setting.

In the copycat problem, the learner does not observe the realized payoffs but knows that the payoff function belongs to a specific set U. The objective is to guarantee performance comparable to the value of the game which is always zero, regardless of the chosen payoff. By leveraging the equivalence mentioned above and viewing the problem through the lens of online optimization, we derive a general solution as Algorithm 1 below.

Algorithm 1 Playing Games with Unobserved Payoffs

input: learning rate $\mu > 0$.
init $u_1(s, a^1, a^2) = 0 \ \ \forall s \in S$ and $a^1, a^2 \in A$.
for $t = 1$ **to** T **do**
  compute $\pi_t^1 = \arg\max_{\pi^1} \min_{\pi^2} \mathbb{E}^{\pi^1,\pi^2}\left[u_t(s, a^1, a^2)\right]$
  play π_t^1 and observe trajectory $(s_{1,t}, a_{1,t}^1, a_{1,t}^2, \ldots, s_{H,t}, a_{H,t}^1, a_{H,t}^2)$.
  $u_{t+1} \leftarrow u_t$.
  for $h = 1$ **to** H **do**
    update $u_{t+1}(s_{h,t}, a^1, a^2) = \begin{cases} u_t(s_{h,t}, a^1, a^2) - \frac{\mu}{2}, & a^1 = a_{h,t}^1, a^2 = a_{h,t}^2, a_{h,t}^1 \neq a_{h,t}^2; \\ u_t(s_{h,t}, a^1, a^2) + \frac{\mu}{2}, & a^1 = a_{h,t}^2, a^2 = a_{h,t}^1, a_{h,t}^1 \neq a_{h,t}^2; \\ u_t(s_{h,t}, a^1, a^2), & \text{otherwise.} \end{cases}$
  end for
  clip $u_{t+1}(s, a^1, a^2)$ entry-wise to $[-1, 1]$ for all $s \in S$, $a^1, a^2 \in A$.
end for

We conceptualize the algorithm as involving a third "virtual player" who observes the interaction between the learner and the adversary. At the beginning of each round t, this virtual player selects a payoff function $u_t \in U$ with the goal of minimizing their cumulative loss. Given a realized trajectory $\tau = (s_1, a_1^1, a_1^2, \ldots, s_H, a_H^1, a_H^2)$, the loss for episode t is defined as the total payoff under the selected function: $\sum_{h=1}^{H} u_t(s_h, a_h^1, a_h^2)$

To optimize this objective, the virtual player employs an Online Gradient Descent algorithm over the set U. At the end of round t, OGD receives a feedback loss tensor $l_t \in \mathbb{R}^{|S| \times n \times n}$. This tensor acts as a sparse indicator of the trajectory: its entries are set to 1 for the H state-action triplets visited in τ and 0 otherwise. In this manner we represent the loss as a linear function of the payoff matrix: $l_t \cdot u_t = \sum_{h=1}^{H} u_t(s_h, a_h^1, a_h^2)$.

Proof Sketch. The proof of the $O(n\sqrt{T|S|H})$ bound is detailed in Section E. The analysis proceeds naturally from the observation that the safety-level strategy ensures a non-negative expected payoff

at each individual state, and consequently, across the entire trajectory of the game. Combining this structural property with the regret bound of OGD establishes the final guarantee. Crucially, implementing this approach for SSG is highly computationally efficient. The per-step complexity remains strictly polynomial: because SSG enforces symmetry independently at each state, restricting the expected further value of any state to zero, both the safety-level policy computation and the projection operate over uncoupled local action spaces, making them trivial to compute. □

5.1 General sum case

We further extend our results to general-sum symmetric games, by considering the normalized "skew-symmetrized" payoff $\hat{u}(s, a^1, a^2) = \frac{1}{2}(u_\star(s, a^1, a^2) - u_\star(s, a^2, a^1))$, such that $\hat{u}(s, a^1, a^2) \in [-1, 1]$. The learner's payoff with respect to this transformed game is similarly bounded:

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H} u_\star(s_{h,t}, a^1_{h,t}, a^2_{h,t}) - u_\star(s_{h,t}, a^2_{h,t}, a^1_{h,t})\right|\right] \leq \sqrt{2T|S|Hn^2}.$$

This implies a powerful result: that even in general-sum SSG, the learner can enforce a symmetric outcome on average regardless of the opponent's strategy. Namely, even if the adversary plays sub-optimally or maliciously, the learner effectively "mirrors" the opponent's success (or failure), ensuring they are never exploited.

Next, we explore learning under the even stricter condition of Assumption 2.

6 Copycat Strategy in HSG

In this section, we prove Theorem 2. We show that the HSG assumption (Assumption 2) imposes a strong structural restriction: once the players choose different actions, their subsequent choices cannot affect the expected remaining payoff. Consequently, after payoff redistribution, the game is equivalent to an SSG that terminates upon the first disagreement between the players.

Just as not every SSG is an MSG, the HSG class is a strictly smaller subset of MSG. While HSGs may still involve complex stochastic transitions, the requirement for symmetry over *history-dependent* policies severely restricts the strategic influence of future actions. We illustrate this distinction with the counterexample in Section H.2.

For a history z_h, we denote by $V^{\pi^1,\pi^2}(z_h)$ the expected cumulative payoff from time step h onward, starting from z_h and following policies π^1 and π^2.

To formalize our analysis, we exploit the ability of history-dependent policies to condition on the entire trajectory. This yields the following structural property:

Lemma 4 (Policy-Independent Continuation After Disagreement)**.** *Consider an HSG and let $z_h \in Z$ be a reachable history containing a disagreement, i.e., $a^1_j \neq a^2_j$ for some $j < h$. Then, for any history-dependent policies $\pi^1, \pi^2, \pi^{1\prime}, \pi^{2\prime}$, the continuation value is identical:*

$$V^{\pi^1,\pi^2}(z_h) = V^{\pi^{1\prime},\pi^{2\prime}}(z_h).$$

Proof. Fix a reachable history z_h containing a disagreement. Choose two history-dependent policies π^1, π^2 that deterministically play the actions specified by z_h at every time step $j < h$, so that $\mathbb{P}^{\pi^1,\pi^2}(z_h) > 0$. Since the history contains a disagreement, swapping these policies makes z_h unreachable:

$$\mathbb{P}^{\pi^2,\pi^1}(z_h) = 0.$$

Now consider another pair of policies $\pi^{1\prime}, \pi^{2\prime}$ that agrees with the original pair outside the continuation of z_h, but follows arbitrary continuation policies after z_h. Applying the HSG assumption to both policy pairs and subtracting the resulting equalities yields

$$\mathbb{P}^{\pi^1,\pi^2}(z_h)\left(V^{\pi^1,\pi^2}(z_h) - V^{\pi^{1\prime},\pi^{2\prime}}(z_h)\right) = 0.$$

Since $\mathbb{P}^{\pi^1,\pi^2}(z_h) > 0$, the continuation values are equal. As the continuation policies were arbitrary, the result follows. □

Combining this structural property with the EPR transformation establishes Theorem 2.

Proof of Theorem 2. Since every HSG is also an MSG, by Lemma 3 the game can be transformed into a value-equivalent SSG. As the transformation preserves the value for all policies, the transformed game also satisfies HSG.

Consider any reachable history z_h containing a disagreement. By Lemma 4, its continuation value is independent of future policies. Since the transformed game is an SSG, playing identical actions at every subsequent step yields zero payoff. Therefore, the continuation value must be zero.

In particular, choosing arbitrary actions a^1, a^2 at z_h and identical actions thereafter gives

$$\hat{u}(s_h, a^1, a^2) = 0.$$

Hence, all payoffs following the first disagreement vanish, and the transformed game can terminate at that point without changing its expected value. □

7 Limitations

We adopt a finite-horizon, episodic formulation to provide a clear unit of interaction and measure learning progress across repeated episodes. The finite horizon also allows us to present Expected Payoff Redistribution through backward induction, making the structural reduction and its analysis explicit.

This formulation does not exclude cyclic dynamics. A physical state may be revisited within an episode; our time-augmented representation assigns each visit to a distinct state–time pair. The resulting partitioned state space is acyclic because time advances, even when the underlying physical process contains cycles.

Our choice is primarily one of presentation and analytical scope. In finite-state discounted infinite-horizon games, backward induction can be replaced by a convergent potential transformation yielding an equivalent per-state symmetric game. Since the learner observes only states and actions, this transformation is needed only for the analysis. The finite-horizon formulation lets us state the learning guarantees without introducing discount-dependent bounds or a separate performance criterion for continuing interaction.

8 Acknowledgments

This project is supported by the Israel Science Foundation (ISF, grant number 2250/22).

References

Jacob Abernethy, Peter L Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 27–46. JMLR Workshop and Conference Proceedings, 2011.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6(4):1–8, 1956.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Near-optimal policy optimization for correlated equilibrium in general-sum markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 3889–3897. PMLR, 2024.

Qiwen Cui, Kaiqing Zhang, and Simon Du. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2651–2652. PMLR, 2023.

Liad Erez, Tal Lancewicki, Uri Sherman, Tomer Koren, and Yishay Mansour. Regret minimization and convergence to equilibria in general-sum markov games. In *International Conference on Machine Learning*, pages 9343–9373. PMLR, 2023.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Michal Feldman, Adam Kalai, and Moshe Tennenholtz. Playing games without observing payoffs. In *ICS*, pages 106–110, 2010.

János Flesch, Thiruvenkatachari Parthasarathy, Frank Thuijsman, and Philippe Uyttendaele. Evolutionary stochastic games. *Dynamic Games and Applications*, 3(2):207–219, 2013.

Elad Hazan. *Introduction to Online Convex Optimization*. MIT Press, 2022.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33:16557–16566, 2020.

Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *International Conference on Machine Learning*, pages 14036–14053. PMLR, 2022.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Nahum Shimkin. An online convex optimization approach to blackwell's approachability. *Journal of Machine Learning Research*, 17(129):1–23, 2016.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In *International conference on machine learning*, pages 10279–10288. PMLR, 2021.

J von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

J van der Wal and Jaap Wessels. On markov games. *Statistica Neerlandica*, 30(2):51–71, 1976.

Yuepeng Yang and Cong Ma. $O(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. *arXiv preprint arXiv:2209.12430*, 2022.

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Satisficing paths and independent multiagent reinforcement learning in stochastic games. *SIAM Journal on Mathematics of Data Science*, 5(3): 745–773, 2023.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

Martin Zinkevich and Tucker R Balch. Symmetry in markov decision processes and its implications for single agent and multiagent learning. In *Proceedings of the Eighteenth International Conference on Machine Learning*, page 632, 2001.

Martin A. Zinkevich. Generalized symmetry in stochastic games. Unpublished manuscript, 2006. URL `https://www.cs.cmu.edu/afs/cs/Web/People/maz/publications/symmetrysg.ps`.

A Copycat in matrix games

Algorithm 2 Copycat in matrix games

Get Learning rate μ
Initialize $u_1(a^1, a^2) = 0 \ \ \forall a^1, a^2 \in A$
for $t = 1$ **to** T **do**
  Compute: $\pi_t^1 = \arg\max_{\pi^1} \min_{\pi^2} \mathbb{E}^{\pi^1,\pi^2}[u_t(\pi^1, \pi^2)]$.
  Play $a_t^1 \sim \pi_t^1$. Observe opponent's action a_t^2.
  Update: $u_{t+1}(a^1, a^2) = \begin{cases} u_t(a^1, a^2) - \frac{\mu}{2}, & a^1 = a_t^1, a^2 = a_t^2, a_t^1 \neq a_t^2; \\ u_t(a^1, a^2) + \frac{\mu}{2}, & a^1 = a_t^2, a^2 = a_t^1, a_t^1 \neq a_t^2; \\ u_t(a^1, a^2), & \text{otherwise} \end{cases}$
  Clip u_{t+1} to $[-1, 1]$
end for

In this section we will introduce an alternative approach to analyzing the copycat problem in symmetric zero-sum matrix games via a reduction to Online Convex Optimization. To that end, we view the algorithm as introducing a third virtual player who observes the actions of the other two players at each round. At time t, this virtual player selects $u_t \in U$, where U represents the set of all skew-symmetric matrices of size $n \times n$. The virtual player's loss, given that the original players chose actions a^1 and a^2 is $u_t(a^1, a^2)$. Like the learner, this virtual player does not observe the true payoff of the game, and updates their plays using only the past plays of the learner and the opponent. The goal of the virtual player is to decrease the virtual payoff in action pairs a^1, a^2 that were played in the past, in order to encourage the players to choose actions a^2, a^1 in the next rounds, which will motivate symmetric play over time.

The virtual player's strategy relies on an online Gradient Descent algorithm over the set of skew-symmetric bounded matrices, defined as

$$U = \{u \in \mathbb{R}^{n \times n} : |u(a^1, a^2)| \leq 1 \text{ and } u(a^1, a^2) = -u(a^2, a^1), \ \forall a^1, a^2 \in A\}.$$

The loss matrices received by OGD in T rounds are $l_1, \cdots, l_T \in \mathbb{R}^{n \times n}$ where each l_t is a matrix whose entries are 0 except for a single entry a_t^1, a_t^2, corresponding to the actions chosen by both players on round t, that equals 1. Note that $\|u\|_F \leq n$ for all $u \in U$ and that $\|l_t\|_F \leq 1$ for all $t = 1, \ldots, T$. As an example for such an algorithm, we can take Online Gradient Descent that has the following regret guarantee (see, e.g., Hazan, 2022):

$$\sum_{t=1}^{T} u_t \cdot l_t - \min_{u \in U} \sum_{t=1}^{T} u \cdot l_t = O(n\sqrt{T}). \tag{1}$$

Given such an online Gradient Descent algorithm, we introduce the learner's strategy in Algorithm 2. On each round t, OGD outputs a skew-symmetric payoff function $u_t \in U$. The learner then selects a safety-level strategy π_t^1 based on u_t. This approach guarantees a non-negative expected payoff w.r.t. u_t, ensuring that:

$$\mathbb{E}\big[u_t(a_t^1, a_t^2)\big] \geq 0, \quad \forall t = 1, \ldots, T, \tag{2}$$

regardless of the strategies $\pi_1^2, \ldots, \pi_T^2$ chosen by the adversary.

Now indeed suppose that the virtual player uses Online Gradient Descent, let $u_\star$ represent the actual payoffs of the game. We bound the learner's payoff as follows:

$$\begin{aligned} \mathbb{E}\left[\left|\sum_{t=1}^{T} u_\star(a_t^1, a_t^2)\right|\right] &= \mathbb{E}\left[-\min_{u \in \{u_\star, -u_\star\}} \sum_{t=1}^{T} u(a_t^1, a_t^2)\right] && (u \text{ is symmetric}) \\ &\leq \mathbb{E}\left[-\min_{u \in U} \sum_{t=1}^{T} u(a_t^1, a_t^2)\right] && (u_\star, -u_\star \in U) \end{aligned}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} u_t(a_t^1, a_t^2) - \min_{u \in U} \sum_{t=1}^{T} u(a_t^1, a_t^2)\right] \quad \text{(Eq. (2))}$$

$$= O(n\sqrt{T}). \quad \text{(Eq. (1))}$$

We have shown this method achieves a bound comparable to that of Feldman et al. [2010]. To complete the analysis, we now show Algorithm 2 runs in polynomial time.

A.1 Computational Complexity

At each round, Algorithm 2 begins with a skew-symmetric matrix u (for example the zero matrix), and updates only a single entry (a_t^1, a_t^2) of the matrix. Consequently, the algorithm must project the updated matrix onto the set of all skew-symmetric matrices. Due to the linear nature of the projection and the fact that only one entry is modified, it suffices to project the updated entry (a_t^1, a_t^2) along with its corresponding symmetric counterpart (a_t^2, a_t^1), which takes $O(1)$ time. As computing a safety-level strategy can be done in $\mathrm{poly}(n)$ time, Algorithm 2 runs in polynomial time in its entirety.

B Lower Bound Analysis

Since the learner does not observe payoffs, they are compelled to play a strategy that is robust against any potential payoff function. This setting is analytically equivalent to a scenario where the adversary selects the worst-case payoff function after the game concludes. We begin by analyzing the fundamental case of a repeated matrix game.

Let $N_T(a^1, a^2)$ denote the number of rounds in which the learner played action a^1 and the adversary played action a^2. We define the discrepancy $D_T(a^1, a^2) := N_T(a^2, a^1) - N_T(a^1, a^2)$.

Consider the following adversarial payoff function, defined in hindsight based on the final action discrepancies:

$$u(a^1, a^2) = \begin{cases} 1, & D_T(a^2, a^1) < 0; \\ -1, & D_T(a^2, a^1) > 0; \\ 0, & \text{otherwise} \end{cases}$$

Observe that for this specific payoff function, the cumulative contribution of any symmetric pair (i, j) satisfies:

$$u(a^1, a^2)N(a^1, a^2) + u(a^2, a^1)N(a^2, a^1) = -\big|D_T(a^1, a^2)\big|$$

Substituting this into the expected payoff guarantee yields:

$$\begin{aligned} \mathbb{E}\left[\left|\sum_{t=1}^{T} u(a_t^1, a_t^2)\right|\right] &= \mathbb{E}\left[\left|\sum_{a^1 < a^2} u(a^1, a^2)N(a^1, a^2) + u(a^2, a^1)N(a^2, a^1)\right|\right] \\ &= \mathbb{E}\left[\left|\sum_{a^1 < a^2} -\big|D_T(a^1, a^2)\big|\right|\right] \\ &= \mathbb{E}\left[\sum_{a^1 < a^2} \big|D_T(a^1, a^2)\big|\right] \end{aligned} \quad (3)$$

Thus, to lower bound the payoff, it suffices to lower bound the expected total discrepancy $\mathbb{E}\big[\sum_{a^1 < a^2} \big|D_T(a^1, a^2)\big|\big]$

B.1 Matrix game

We establish a lower bound of $\Omega(n\sqrt{T})$ on the expected cumulative regret. We assume that the number of actions is fixed and satisfies $n > 2$, that the time horizon T is sufficiently large, and that the adversary plays a **uniform strategy**, choosing each action with probability $1/n$ independently at every step.

Let $N_t(a)$ denote the number of times the learner plays action a up to time t. We analyze two regimes based on the learner's action frequencies.

Case 1: The Unbalanced Regime Suppose there exists an action a^1 such that the learner's expected count deviates significantly from uniformity:

$$\mathbb{E}\left[\left|N_T(a^1) - \frac{T}{n}\right|\right] \geq 2n\sqrt{T}.$$

To lower bound the regret, we decompose the discrepancy using the Reverse Triangle Inequality:

$$\mathbb{E}\big[\big|D_T(a^2, a^1)\big|\big] \geq \underbrace{\frac{1}{n}\mathbb{E}\big[\big|N_T(a^1) - N_T(a^2)\big|\big]}_{\text{Signal}} - \underbrace{\mathbb{E}\left[\left|N_T(a^1, a^2) - N_T(a^2, a^1) - \frac{1}{n}N_T(a^1) + \frac{1}{n}N_T(a^2)\right|\right]}_{\text{Noise}}.$$

First, we bound the **Total Signal**. Using the inequality $\sum_{a^2 \in A} |N_T(a^1) - N_T(a^2)| \geq n|N_T(a^1) - \frac{T}{n}|$, we obtain:

$$\sum_{a^2 \neq a^1} \frac{1}{n}\mathbb{E}\big[\big|N_T(a^1) - N_T(a^2)\big|\big] \geq \mathbb{E}\left[\left|N_T(a^1) - \frac{T}{n}\right|\right] \geq 2n\sqrt{T}. \tag{4}$$

Next, we bound the **Total Noise**. We define the noise term $\mathcal{E}_t$ as:

$$\mathcal{E}_t \coloneqq \mathbb{I}\{a_t^1 = a^1, a_t^2 = a^2\} - \mathbb{I}(a_t^1 = a^2, a_t^2 = a^1) - \frac{1}{n}\mathbb{I}(a_t^1 = a^1) + \frac{1}{n}\mathbb{I}(a_t^1 = a^2).$$

The total noise corresponds to the magnitude of the sum of these variables. By applying Jensen's inequality and bounding the variance, we have:

$$\mathbb{E}\left[\left|\sum_{t=1}^{T} \mathcal{E}_t\right|\right] \leq \sqrt{\mathrm{Var}\left(\sum_{t=1}^{T} \mathcal{E}_t\right)} \leq \sqrt{\frac{1}{n}\mathbb{E}[N_T(a^1) + N_T(a^2)]}.$$

Summing this bound over all $a^2 \neq a^1$ and applying the Cauchy–Schwarz inequality to the sum of $n - 1$ terms yields:

$$\sum_{a^2 \neq a^1} \mathbb{E}\left[\left|\sum_{t=1}^{T} \mathcal{E}_t\right|\right] \leq \sqrt{(n-1)\sum_{a^2 \neq a^1} \frac{1}{n}\mathbb{E}[N_T(a^1) + N_T(a^2)]} \leq \sqrt{nT}.$$

Conclusion. Combining the signal bound with the noise bound above yields

$$\mathbb{E}\left[\sum_{a^1 < a^2} \big|D_T(a^2, a^1)\big|\right] \geq \sum_{a^2 \neq a^1} \mathbb{E}\big[\big|D_T(a^2, a^1)\big|\big] \geq 2n\sqrt{T} - \sqrt{nT} \geq n\sqrt{T},$$

which holds for all sufficiently large T.

Case 2: The Balanced Regime Suppose that for all actions $a \in A$, the counts are concentrated near the uniform allocation:

$$\mathbb{E}\left[\left|N_T(a) - \frac{T}{n}\right|\right] < 2n\sqrt{T}.$$

We note that by **Jensen's Inequality**:

$$\left|\mathbb{E}[N_T(a)] - \frac{T}{n}\right| \leq \mathbb{E}\left[\left|N_T(a) - \frac{T}{n}\right|\right] < 2n\sqrt{T}.$$

We lower-bound using the specific payoff of a Rock-Paper-Scissors (RPS) game on the cycle $\Delta = \{a^1, a^2, a^3\}$. Define the payoff function $u(a_t^1, a_t^2)$ as 1 if the pair (a_t^1, a_t^2) is a winning edge for the learner in Δ, -1 if it is a losing edge, and 0 otherwise.

Let $X \coloneqq \sum_{t=1}^{T} u(a_t^1, a_t^2)$. By the triangle inequality, the total discrepancy on the cycle upper-bounds the magnitude of this cumulative payoff:

$$\sum_{\substack{a^1 \le a^2 \\ a^1, a^2 \in \Delta}} \big|D_T(a^1, a^2)\big| \ge |X|.$$

To lower-bound $\mathbb{E}[|X|]$, we employ the Fourth Moment Method (a consequence of Hölder's inequality):

$$\mathbb{E}[|X|] \ge \frac{(\mathbb{E}[X^2])^{3/2}}{(\mathbb{E}[X^4])^{1/2}}.$$

The Second Moment. The sequence $(u(a_t^1, a_t^2))_{t=1}^T$ forms a martingale difference sequence. The conditional variance at step t is non-zero only if the learner plays an action $i_t \in \Delta$ and the adversary (playing uniformly) hits a neighbor in the cycle (with probability $2/n$). Summing over all time steps:

$$\mathbb{E}[X^2] \;=\; \sum_{t=1}^{T} \mathbb{E}\big[\mathbb{E}\big[u(a_t^1, a_t^2)^2 \mid a_1^1, a_1^2 \ldots a_{t-1}^1, a_{t-1}^2, a_t^1\big]\big] = \sum_{t=1}^{T} \mathbb{E}\left[\frac{2}{n}\mathbb{I}_{\{a_t^1 \in \Delta\}}\right] = \frac{2}{n}\mathbb{E}[N_\Delta],$$

where $N_\Delta = \sum_{a \in \Delta} N_T(a)$ denotes the total number of times the learner plays in the cycle.

The Fourth Moment. Let $Y_t = u(a_t^1, a_t^2)$ and $q_t = \frac{2}{n}\mathbf{1}\{a_t^1 \in \Delta\}$. Since the opponent plays uniformly, $\mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}, a_t^1] = q_t$. Thus, $Z_t = Y_t^2 - q_t$ is a martingale difference sequence, and

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} Z_t\right)^2\right] \le \sum_{t=1}^{T} \mathbb{E}[q_t] = \frac{2}{n}\mathbb{E}[N_\Delta].$$

By the Burkholder–Davis–Gundy inequality,

$$\begin{aligned}
\mathbb{E}\big[X^4\big] &\le C\mathbb{E}\left[\left(\sum_{t=1}^{T} Y_t^2\right)^2\right] \\
&= C\mathbb{E}\left[\left(\sum_{t=1}^{T} (Z_t + q_t)\right)^2\right] \\
&\le 2C\mathbb{E}\left[\left(\sum_{t=1}^{T} Z_t\right)^2\right] + 2C\mathbb{E}\left[\left(\sum_{t=1}^{T} q_t\right)^2\right] \\
&\le \frac{4C}{n}\mathbb{E}[N_\Delta] + \frac{8C}{n^2}\mathbb{E}\big[N_\Delta^2\big].
\end{aligned}$$

(Note: The last inequality follows from the martingale difference property of Z_t, together with $(a+b)^2 \le 2a^2 + 2b^2$.)

Since N_Δ scales linearly with T (specifically $\mathbb{E}[N_\Delta] \propto T/n$), the quadratic term $\mathbb{E}[N_\Delta^2]$ grows as T^2, dominating the linear term $\mathbb{E}[N_\Delta]$. Thus, for sufficiently large T, we can absorb the linear term into the quadratic term:

$$\mathbb{E}\big[X^4\big] \le \frac{16C}{n^2}\,\mathbb{E}\big[N_\Delta^2\big].$$

Bounding the Second Moment. We show that the second moment $\mathbb{E}[N_\Delta^2]$ is dominated by the squared mean. Since N_Δ and $\mathbb{E}[N_\Delta]$ are both restricted to the interval $[0, T]$, their distance is at most T. Thus, almost surely:

$$(N_\Delta - \mathbb{E}[N_\Delta])^2 \le T|N_\Delta - \mathbb{E}[N_\Delta]|.$$

Taking expectations and applying the regime condition $\mathbb{E}[|N(a) - T/n| < 2n\sqrt{T}]$:

$$\mathrm{Var}[N_\Delta] = \mathbb{E}\big[(N_\Delta - \mathbb{E}[N_\Delta])^2\big] \le T \sum_{a \in \Delta} \mathbb{E}[|N_a - \mathbb{E}[N_a]|]$$

$$\leq \sum_{a \in \Delta} 2T\mathbb{E}\left[\left|N_T(a) - \frac{T}{n}\right|\right] \leq 12nT\sqrt{T}$$

The variance scales as $O(T^{1.5})$, which is asymptotically negligible compared to $\mathbb{E}[N_\Delta]^2 \propto T^2$. Implying that for large enough $T > n^6$ we will have

$$\mathbb{E}[N_\Delta^2] = \mathrm{Var}[N_\Delta] + \mathbb{E}[N_\Delta]^2 \leq 3\mathbb{E}[N_\Delta]^2$$

Completing the Bound. We now substitute these moments into the Hölder's ratio. Squaring both sides implies:

$$\mathbb{E}[|X|]^2 \geq \frac{\left(\mathbb{E}[X^2]\right)^3}{\mathbb{E}[X^4]} \geq \frac{(\frac{2}{n}\mathbb{E}[N_\Delta])^3}{\frac{16C}{n^2}\,\mathbb{E}[N_\Delta^2]}.$$

Substituting the second moment bound into the denominator:

$$\mathbb{E}[|X|]^2 \geq \frac{(\frac{2}{n}\mathbb{E}[N_\Delta])^3}{\frac{16C}{n^2} \cdot 3\,\mathbb{E}[N_\Delta]^2} \geq \frac{1}{6Cn}\mathbb{E}[N_\Delta] \geq \frac{1}{2Cn}(\frac{T}{n} - 2n\sqrt{T}) = \Omega\left(\frac{T}{n^2}\right).$$

Finally, taking the square root implies:

$$\mathbb{E}[|X|] = \Omega\left(\frac{\sqrt{T}}{n}\right).$$

Aggregation. Summing over all $\binom{n}{3}$ cycles and dividing by the overcounting factor $(n-2)$ (since each edge is part of $n-2$ triangles):

$$\mathbb{E}\left[\sum_{i<j}|D_T(i,j)|\right] \;\geq\; \frac{1}{n-2}\sum_{\{i,j,k\}}\mathbb{E}[|X_{i,j,k}|] \;=\; \frac{1}{n-2}\cdot\binom{n}{3}\Omega(\frac{\sqrt{T}}{n}) \;=\; \Omega(n\sqrt{T}).$$

B.2 SSG games

We extend the lower bound to the Markovian setting by constructing a specific SSG instance with a constant $|S_h|$ states for any step h and episodes of length H and a starting state s_1 with a zero payoff matrix. Consider a game where the transition dynamics are uniform and independent of the current state or actions. Specifically, for any state s, step h, and action pair (a^1, a^2):

$$P(s'_{h+1} \mid s_h, a^1, a^2) = \frac{1}{|S_{h+1}|}$$

In this construction, the expected visitation frequency of any state is a random variable determined solely by the transition, independent of the players' strategies. Let T_s denote the random number of times state s is visited over the entire game horizon of T episodes. The expected number of visits is $\mathbb{E}[T_s] = \frac{T}{|S_h|}$.

Since the states are decoupled by the uniform transitions, the game effectively decomposes into $|S|$ independent repeated matrix games, where the effective time horizon for state s is T_s. Applying the reduction from Eq. (3) locally to each state:

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H}u(s_{h,t}, a_t^1, a_t^2)\right|\right] = \mathbb{E}\left[\sum_{h=2}^{H}\sum_{s\in S_h}\sum_{a^1<a^2}\left|D_{T_s}(s, a^1, a^2)\right|\right]$$

We assume that $T \geq Cn^6|S|$ for a sufficiently large constant C. For each $s \in S_h$, $h \geq 2$, we have $T_s \sim \mathrm{Binomial}(T, 1/|S_h|)$. By a Chernoff bound,

$$\Pr\left(T_s \geq \frac{T}{2|S_h|}\right) \geq \frac{1}{2}.$$

Conditioning on the visitation sequence and applying the matrix-game lower bound yields, for every such state,

$$\mathbb{E}\left[\sum_{a^1<a^2}\left|D_{T_s}(s,a^1,a^2)\right|\right]=\Omega\left(n\sqrt{\frac{T}{|S_h|}}\right).$$

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H}u(s_{h,t},a_t^1,a_t^2)\right|\right]&=\mathbb{E}\left[\sum_{h=2}^{H}\sum_{s\in S_h}\sum_{a^1<a^2}\left|D_{T_s}(s,a^1,a^2)\right|\right]\\&=\sum_{h=2}^{H}\sum_{s\in S_h}\mathbb{E}\left[\sum_{a^1<a^2}\left|D_{T_s}(s,a^1,a^2)\right|\right]\\&\geq|S_h|(H-1)\cdot\Omega\left(n\sqrt{\frac{T}{|S_h|}}\right)\\&=\Omega\left(n\sqrt{H|S|T}\right).\qquad(|S|=1+(H-1)|S_h|)\end{aligned}$$

C Matrix game as Blackwell's Approachability

We first define the main concepts of approachability theory. A Blackwell instance is a tuple $(\mathcal{X},\mathcal{Y},L,U)$ such that $\mathcal{X}\subset\mathbb{R}^n$ is the action set of the learner, $\mathcal{Y}\subset\mathbb{R}^m$ is the action set of the adversary. Both action sets are convex and compact. The set $U\subset\mathbb{R}^d$ is convex and closed, and the vector valued function $L:\mathcal{X}\times\mathcal{Y}\mapsto\mathbb{R}^d$ is bilinear.

Definition 2 (Approachable set)**.** *Given a Blackwell instance $(\mathcal{X},\mathcal{Y},L,U)$, we say that U is* approachable *if there exists an algorithm* ALG *that selects points $x^1,\dots,x^T\in\mathcal{X}$ such that, for any sequence $y^1,y^2,\dots y^T\in\mathcal{Y}$, we have:*

$$\operatorname{dist}\left(\frac{1}{T}\sum_{t=1}^{T}L(x^t,y^t),U\right)\to 0\quad as\quad T\to\infty,$$

where $x^t\leftarrow\mathrm{ALG}(y^1,y^2,\dots y^{t-1})$.

Finally, we state Blackwell's Approachability Theorem.

Theorem 4. *(Approachability Theorem; Blackwell, 1956) For any Blackwell instance $(\mathcal{X},\mathcal{Y},L,U)$, U is approachable if and only if $\forall y\in\mathcal{Y}\ \exists x\in\mathcal{X}\ :L(x,y)\in U$.*

C.1 Formulation of the copycat problem as Blackwell's Approachability

We next formulate the copycat problem as a Blackwell's Approachability problem. We define a Blackwell instance $(\Delta A,\Delta A,L,U)$, such that $A=\{1,\dots,n\}$ is the action set of both players. The target set U is the space of all symmetric matrices:

$$U=\{u\in\mathbb{R}^{n\times n}:u(a^1,a^2)=u(a^2,a^1),\ \forall a^1,a^2\in A)\}.$$

We define $L(x,y)=xy^\top$, where $x,y\in\Delta A$. For pure actions $i,j\in A$, we write $L(i,j)=e_ie_j^\top$, where e_i is the i-th standard basis vector.

Lemma 5. *The copycat problem is approachable.*

Proof. For each $y\in\Delta A$ we choose $x=y$ Then $L(y,y)=yy^\top$ is symmetric and therefore belongs to U. □

Lastly, consider a repeated symmetric matrix game with payoff $u_\star$. Then, since the problem is approachable:

$$\left|\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}u_\star(a_t^1,a_t^2)\right]\right|=\left|u_\star\cdot\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}L(a_t^1,a_t^2)\right]\right|\to 0.$$

D Proof of Lemma 1

Proof. We look at $\pi^1, \hat{\pi}^1, \pi^2, \hat{\pi}^2 \in \Pi_M$ under the special case where $\pi^1 = \hat{\pi}^1$ and $\pi^2 = \hat{\pi}^2$ for all h except H and for all states except $s \in S_H$. In this case:

$$\begin{aligned}
&V^{\pi^1,\pi^2}(s_1) - V^{\hat{\pi}^1,\hat{\pi}^2}(s_1) \\
&\quad = \mathbb{P}^{\pi^1,\pi^2}[s_H = s]\Big(\mathbb{E}^{\pi^1,\pi^2}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big] - \mathbb{E}^{\hat{\pi}^1,\hat{\pi}^2}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big]\Big).
\end{aligned}$$

Assumption 1 implies:

$$\begin{aligned}
&V^{\pi^1,\pi^2}(s_1) - V^{\hat{\pi}^1,\hat{\pi}^2}(s_1) + V^{\pi^2,\pi^1}(s_1) - V^{\hat{\pi}^2,\hat{\pi}^1}(s_1) \\
&\quad = (\underbrace{V^{\pi^1,\pi^2}(s_1) + V^{\pi^2,\pi^1}(s_1)}_{=0}) - (\underbrace{V^{\hat{\pi}^1,\hat{\pi}^2}(s_1) + V^{\hat{\pi}^2,\hat{\pi}^1}(s_1)}_{=0}) \\
&\quad = 0.
\end{aligned}$$

We combine both observations to get:

$$\begin{aligned}
&\mathbb{P}^{\pi^1,\pi^2}[s_H = s](\mathbb{E}^{\pi^1,\pi^2}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big] - \mathbb{E}^{\hat{\pi}^1,\hat{\pi}^2}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big]) \\
&\quad = -\mathbb{P}^{\pi^2,\pi^1}[s_H = s](\mathbb{E}^{\pi^2,\pi^1}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big] - \mathbb{E}^{\hat{\pi}^2,\hat{\pi}^1}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big]).
\end{aligned} \tag{5}$$

Now we set π^1, π^2 such that they will always choose action 1 in $s \in S_H$. Formally, $\mathbb{P}^{\pi^1}[a_H = 1 \mid s_H = s] = \mathbb{P}^{\pi^2}[a_H = 1 \mid s_H = s] = 1$. Plugging these policies into Eq. (5) gets:

$$\begin{aligned}
&\mathbb{P}^{\pi^1,\pi^2}[s_H = s]\Big(u(s,1,1) - \mathbb{E}^{\hat{\pi}^1,\hat{\pi}^2}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big]\Big) = \\
&\qquad - \mathbb{P}^{\pi^2,\pi^1}[s_H = s]\Big(u(s,1,1) - \mathbb{E}^{\hat{\pi}^2,\hat{\pi}^1}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big]\Big).
\end{aligned} \tag{6}$$

Next we choose π^1, π^2 that choose actions uniformly at random everywhere except in $s \in S_H$. Formally: $\mathbb{P}^{\pi^i}[a = a_h \mid s_h] = \frac{1}{|A|}$ for all $h \in [H]$, $s_h \in S_h \setminus \{s\}$ and $i = 1, 2$. By the choice of our policies, we have that $\mathbb{P}^{\pi^1,\pi^2}[s_H = s] = \mathbb{P}^{\pi^2,\pi^1}[s_H = s] > 0$, from which

$$u(s,1,1) + u(s,1,1) = \mathbb{E}^{\hat{\pi}^1,\hat{\pi}^2}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big] + \mathbb{E}^{\hat{\pi}^2,\hat{\pi}^1}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big],$$

Taking $\hat{\pi}^2$ and $\hat{\pi}^1$ to be deterministic proves Item 1.

Now we rearrange Eq. (6):

$$\begin{aligned}
&\mathbb{P}^{\pi^1,\pi^2}[s_H = s]u(s, a^1, a^2) + \mathbb{P}^{\pi^2,\pi^1}[s_H = s]u(s, a^2, a^1) = \\
&\qquad \mathbb{P}^{\pi^1,\pi^2}[s_H = s]\; u(s,1,1) + \mathbb{P}^{\pi^2,\pi^1}[s_H = s]\; u(s,1,1)
\end{aligned} \tag{7}$$

For simplicity we will define:

$$C(u, \pi^1, \pi^2, h) := \mathbb{E}^{\pi^1,\pi^2}\left[\sum_{h'=1}^{h} u(s_{h'}, a_{h'}^1, a_{h'}^2)\right] + \mathbb{E}^{\pi^2,\pi^1}\left[\sum_{h'=1}^{h} u(s_{h'}, a_{h'}^1, a_{h'}^2)\right].$$

We begin our proof by considering the payoff at step H for policies $\pi_1, \pi_2 \in \Pi_M$:

$$\begin{aligned}
C(u, \pi_1, \pi_2, H) &= C(u, \pi_1, \pi_2, H-1) + \mathbb{E}^{\pi_1,\pi_2}\big[u(s_H, a_H^1, a_H^2)\big] + \mathbb{E}^{\pi_2,\pi_1}\big[u(s_H, a_H^1, a_H^2)\big] \\
&= C(u, \pi_1, \pi_2, H-1) + \sum_{s \in S_H} \mathbb{P}^{\pi^1,\pi^2}[s_H = s]\mathbb{E}^{\pi_1,\pi_2}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big] \\
&\qquad + \sum_{s \in S_H} \mathbb{P}^{\pi^2,\pi^1}[s_H = s]\mathbb{E}^{\pi_2,\pi_1}\big[u(s, a_H^1, a_H^2) \mid s_H = s\big] \\
&= C(u, \pi_1, \pi_2, H-1) + \sum_{s \in S_H} \mathbb{P}^{\pi^1,\pi^2}[s_H = s]u(s,1,1) \\
&\qquad + \sum_{s \in S_H} \mathbb{P}^{\pi^2,\pi^1}[s_H = s]u(s,1,1). \qquad \text{(Eq. (6))}
\end{aligned}$$

Expanding the last two summands using the Markov property:

$$\sum_{s\in S_H} \mathbb{P}^{\pi_1,\pi_2}[s_H = s]u(s,1,1) + \sum_{s\in S_H} \mathbb{P}^{\pi_2,\pi_1}[s_H = s]u(s,1,1)$$
$$= \mathbb{E}^{\pi_1,\pi_2}\left[\sum_{s\in S} P(s \mid s_{H-1}, a^1_{H-1}, a^2_{H-1})u(s,1,1)\right]$$
$$+ \mathbb{E}^{\pi_2,\pi_1}\left[\sum_{s\in S} P(s \mid s_{H-1}, a^1_{H-1}, a^2_{H-1})u(s,1,1)\right].$$

Finally, substituting this back into the original expression for the payoff:

$$C(u,\pi_1,\pi_2,H) = C(\hat{u},\pi_1,\pi_2,H-1),$$

We can clearly see that the left side must satisfy Assumption 1, hence the right side must as well, completing our proof.

□

E Strategy: Learn on Each State Independently

At first glance, generalizing from matrix games to Markov games suggests that the adversary might exploit the transition dynamics to gain an advantage. Since players' actions influence future state visitations, one might assume a holistic approach accounting for long-term planning is required. However, we demonstrate that a local Copycat strategy is surprisingly robust to unknown dynamics. We show that for SSG, the learner can effectively decouple the transition dynamics from the immediate payoffs. By simply treating each state as an independent matrix game, the learner can ignore the transition function P entirely while still guaranteeing sublinear regret, avoiding any exponential scaling with the game's parameters.

The learner's strategy is to select a safety-level strategy π^1_t at each state, treating u_t as the payoff function. As $u_t(s,\cdot,\cdot)$ is skew-symmetric, the value of the game is zero, guaranteeing a non-negative expected payoff for the learner:

$$\mathbb{E}^{\pi^1_t,\pi^2_t}\left[\sum_{h=1}^{H} u_t(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right] \geq \sum_{h=1}^{H} \mathbb{E}^{\pi^1_t,\pi^2_t}\left[u_t(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right] \geq 0, \tag{8}$$

regardless of the strategy π^2_t chosen by the adversary.

Note that $\|u\|_F \leq n\sqrt{|S|}$ for all $u \in U$ and that $\|l_t\|_F \leq \sqrt{H}$ for all $t = 1,\ldots,T$. Hence OGD yields:

$$\sum_{t=1}^{T} u_t \cdot l_t - \min_{u\in U} \sum_{t=1}^{T} u \cdot l_t = O(n\sqrt{T|S|H}). \tag{9}$$

Now indeed suppose that the virtual player uses OGD. Let $u_\star$ represent the actual payoffs of the game. We bound the learner's playoff as follows:

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H} u(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right|\right] = \mathbb{E}\left[-\min_{u\in\{u_\star,-u_\star\}} \sum_{t=1}^{T}\sum_{h=1}^{H} u(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right] \quad \text{(Assumption 3)}$$
$$\leq \mathbb{E}\left[-\min_{u\in U} \sum_{t=1}^{T}\sum_{h=1}^{H} u(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right] \quad (u_\star, -u_\star \in U)$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=1}^{H} u_t(s_{h,t}, a^1_{h,t}, a^2_{h,t}) - \min_{u\in U} \sum_{t=1}^{T}\sum_{h=1}^{H} u(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right] \quad \text{(Eq. (8))}$$

$$= O(n\sqrt{T|S|H}). \qquad \text{(Eq. (9))}$$

Furthermore, we claim that Algorithm 1 runs in polynomial time. Indeed, computing a safety-level strategy takes polynomial time [Wal and Wessels, 1976], and the rest of the steps of the algorithm also run in polynomial time. It remains to show that OGD over U runs in polynomial time, for which it suffices to explain how to perform efficient Euclidean projection onto the set U (see Section 2.2). As the set U is defined such that the payoff matrix in each state is skew-symmetric, projection on this set can be done by projecting the payoff in each state onto the space of skew-symmetric matrices separately, which runs in linear time (see Section A.1). Thus projections on U take $O(|S|n^2)$ time in total. We note that the approach presented here serves as a more structural formulation of the basic Copycat algorithm. Furthermore, we show in Section F that the standard Copycat algorithm can also be generalized to solve SSG environments with similar theoretical guarantees.

Connection to cumulative policy values. For any fixed payoff function u, the difference between the realized return in episode t and $V^{\pi_t^1,\pi_t^2}(s_1)$ has conditional expectation zero, given the preceding history and the policies selected for that episode. These differences are therefore martingale differences. Since each episode's return lies in $[-H, H]$, its conditional variance is at most H^2. Thus, by Cauchy–Schwarz and the orthogonality of martingale differences,

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\left(\sum_{h=1}^{H} u(s_{h,t}, a_{h,t}^1, a_{h,t}^2) - V^{\pi_t^1,\pi_t^2}(s_1)\right)\right|\right]$$
$$\leq \sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left(\sum_{h=1}^{H} u(s_{h,t}, a_{h,t}^1, a_{h,t}^2) - V^{\pi_t^1,\pi_t^2}(s_1)\right)^2\right]} \leq H\sqrt{T}.$$

Consequently, the triangle inequality yields

$$\mathbb{E}\left|\sum_{t=1}^{T} V^{\pi_t^1,\pi_t^2}(s_1)\right| \leq \mathbb{E}\left|\sum_{t=1}^{T}\sum_{h=1}^{H} u(s_{h,t}, a_{h,t}^1, a_{h,t}^2)\right| + H\sqrt{T}.$$

Since $|S| \geq H$ in the layered state space, this preserves the $O(n\sqrt{T|S|H})$ upper bound.

F Strategy: Copycat on Each State

At first glance, generalizing from matrix games to Markov games suggests that the adversary might exploit the transition dynamics to gain an advantage. Since players' actions influence future state visitations, one might assume a holistic approach accounting for long-term planning is required. However, we demonstrate that the Copycat strategy is surprisingly robust to unknown dynamics. We show that for SSG, the learner can effectively decouple the transition dynamics from the immediate payoffs. By simply treating each state as an independent matrix game and applying the Copycat strategy locally, the learner can ignore the transition function P entirely while still guaranteeing sublinear regret, avoiding any exponential scaling with the game's parameters.

To formalize the strategy, we first define the count matrix $N_t \in \mathbb{Z}^{n\times n}$, where $N_t(a^1, a^2)$ denotes the number of times the joint action (a^1, a^2), the learner playing a^1 and the adversary playing a^2, has occurred up to time t (noninclusive). We further define the **discrepancy matrix** D_t as the difference between the counts of symmetric action pairs: $D_t(a^1, a^2) \coloneqq N_t(a^2, a^1) - N_t(a^1, a^2)$.

We analyze the strategy where the learner plays the original Copycat algorithm of Feldman et al. [2010] on each state $s \in S$ independently. Let $u_\star$ denote the true (unobserved) payoff of the game. We define the cumulative payoff α as: $\alpha = \sum_{t=1}^{T}\sum_{h=1}^{H} u_\star(s_{h,t}, a_{h,t}^1, a_{h,t}^2)$. Using the fact that the magnitude of the cumulative payoff is bounded by the final discrepancy of the action counts. Specifically: $|\alpha| \leq \sum_{h=1}^{H}\sum_{s\in S_h}\sum_{a^1,a^2\in A}\frac{1}{2}\left|D_{T+1}(s, a^1, a^2)\right|$ where $D_{T+1}(s, a^1, a^2)$ is the final entry in the discrepancy matrix for state s after T episodes.

and,by Cauchy-Schwartz:

$$(\mathbb{E}(|\alpha|))^2 \leq \mathbb{E}\left[\alpha^2\right] \leq \mathbb{E}\left[\left(\sum_{h=1}^{H}\sum_{s\in S_h}\sum_{a^1,a^2\in A}\frac{1}{2}\left|D_{T+1}(s, a^1, a^2)\right|\right)^2\right]$$

$$\leq \mathbb{E}\left[\frac{|S|n^2}{4}\sum_{h=1}^{H}\sum_{s\in S_h}\sum_{a^1,a^2\in A}(D_{T+1}(s,a^1,a^2))^2\right]$$

To analyze the growth of this term, we define the potential function β_t^h representing the sum of squared discrepancies at time t for states at horizon h: $\beta_t^h = \sum_{s\in S_h}\sum_{a^1,a^2\in A}(D_t(s,a^1,a^2))^2$.

We analyze the single-step difference $\beta_{t+1}^h - \beta_t^h$. Note that in each episode t, exactly one state s_h is visited at each step h. For the unvisited states, the discrepancy does not change. For the visited state $s = s_h$, the change is bounded as follows:

$$\begin{aligned}\beta_{t+1}^h - \beta_t^h &= \begin{cases}0 & \text{if } a_{h,t}^1 = a_{h,t}^2;\\ 2 - 4D_t(s_{h,t}, a_{h,t}^1, a_{h,t}^2) & \text{otherwise.}\end{cases}\\ &\leq 2 - 4D_t(s_{h,t}, a_{h,t}^1, a_{h,t}^2).\end{aligned}$$

Summing over the horizon H:

$$\sum_{h=1}^{H}\beta_{t+1}^h - \beta_t^h \leq \sum_{h=1}^{H} 2 - 4D_t(s_{h,t}, a_{h,t}^1, a_{h,t}^2) = 2H - 4\sum_{h=1}^{H} D_t(s_{h,t}, a_{h,t}^1, a_{h,t}^2).$$

Crucially, the Copycat strategy computes the max-min strategy for the virtual game defined by $D_t(s,\cdot,\cdot)$. This guarantees that the expected "virtual payoff" is non-negative at every step: $\mathbb{E}[D_t(s_h, a_t^1, a_t^2)] \geq 0$. By linearity of expectation: $\mathbb{E}[\sum_{h=1}^{H} D_t(s_{h,t}, a_{h,t}^1, a_{h,t}^2)] = \sum_{h=1}^{H}\mathbb{E}[D_t(s_{h,t}, a_{h,t}^1, a_{h,t}^2)] \geq 0$.

Substituting this back into the potential function inequality, we obtain: $\mathbb{E}[\sum_{h=1}^{H}\beta_{t+1}^h - \beta_t^h] \leq 2H$.

Since $\beta_1^h = 0$ (assuming zero initial discrepancy), telescoping the sum over T yields $\mathbb{E}[\sum_{h=1}^{H}\beta_{T+1}^h] \leq 2HT$. Finally, substituting this bound into our expression for the squared expected payoff:

$$(\mathbb{E}(|\alpha|))^2 \leq \mathbb{E}\left[\frac{|S|n^2}{4}\sum_{h=1}^{H}\beta_{T+1}^h\right] \leq \frac{T|S|Hn^2}{2}.$$

Taking the square root yields the desired bound. This result confirms that in the SSG setting, the adversary's ability to influence the trajectory yields minimal advantage, as they cannot force a worse regret than what is incurred by solving the local matrix games.

G Robustness to Approximate Symmetry (ϵ-Stability)

A natural question is whether our structural reductions and learning guarantees are robust to small deviations from perfect symmetry. If one attempts to push ϵ-inequalities directly through the recursive peeling of the EPR transformation, the exact equalities required for the structural collapse appear to break, as local approximation errors compound iteratively over the horizon.However, we can bypass this algebraic fragility by viewing the problem from a different perspective: reframing the ϵ-approximation as a global model misspecification. Instead of defining the relaxation purely through local ϵ-inequalities over the value function, we define it as the distance to the nearest exact game. Let U be the space of exact MSG (or HSG) payoffs. We assume the true payoff $u^\star$ is ϵ-close to some exact game $u \in U$, such that:

$$||u^\star - u||_1 \leq \epsilon \tag{10}$$

Note that these two definitions are essentially interchangeable, merely yielding different scalings of ϵ. We now prove the following:

Proposition 2. *If the true game $u^\star$ is ϵ-close to an exact symmetric game $u \in U$, deploying our algorithm guarantees that the expected cumulative regret degrades gracefully, bounded by $O(R(u) + \epsilon T)$, where $R(u)$ is the regret of the game under one of our propositions. Furthermore, the learner cannot escape the $\Omega(\epsilon T)$ factor.*

Proof. When the learner deploys our algorithm, they can guarantee that for any $u \in U$, Regret(u) is bounded by the relevant theorem (3, 1, 2).

Lets assume $\hat{u} = \arg\min_{u \in U} ||u^\star - u||_1$ and define $\Delta u = u^\star - \hat{u}$, then:

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H} u^\star(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right|\right] &= \mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H} \hat{u}(s_{h,t}, a^1_{h,t}, a^2_{h,t}) + \Delta u(s_h, a^1, a^2)\right|\right] \\
&\leq \mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H} \hat{u}(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right|\right] + \mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H} \Delta u(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right|\right] \\
&\qquad\qquad \text{(triangle inequality)} \\
&\leq \mathbb{E}\left[\left|\sum_{t=1}^{T}\sum_{h=1}^{H} \hat{u}(s_{h,t}, a^1_{h,t}, a^2_{h,t})\right|\right] + T\epsilon \qquad \text{(Eq. (10))} \\
&\leq R(\hat{u}) + T\epsilon
\end{aligned}$$

Crucially, because our setting assumes the learner cannot observe payoffs, they remain completely blind to the specific direction and structure of this ϵ-perturbation. From a worst-case guarantee perspective, this is mathematically equivalent to an adversary selecting the perturbed payoff function in hindsight. To illustrate why this $\Omega(\epsilon T)$ bound is tight and cannot be improved by the learner, let us consider a fully symmetric SSG with the payoff function u where deviating in s_1 incurs a severe penalty $u(s_1, a^1 \neq 1, a^2 = 1) = -1$. Now, we choose the following ϵ-perturbation:

$$u^\star = \begin{cases} -\epsilon & \text{if } (s, a^1, a^2) = (s_1, 1, 1) \\ u & \text{otherwise} \end{cases}$$

Consider an adversary that always plays action 1 at state s_1. The optimal policy for the learner is to play action 1. By doing so, they inevitably absorb $-\epsilon$ penalty, which is still strictly better than playing any other action and triggering the catastrophic -1 penalty. Consequently,

$$R(u^\star) \geq \epsilon T.$$

Therefore, the additive ϵT term is unavoidable.

□

H Examples

H.1 SSG but not MSG

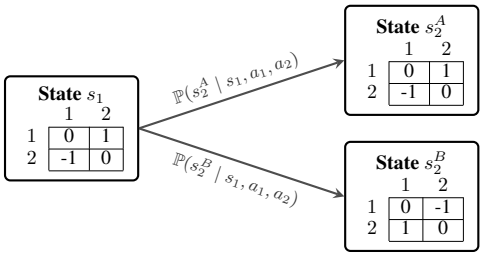


Figure 1: Example: SSG but not MSG game.

In this example, if $P(s_2^A \mid s_1, 1, 2) = P(s_2^B \mid s_1, 2, 1) = 1$, consider the deterministic strategies:

$$\pi^1(s) = \begin{cases} 1, & s = s_1 \\ 1, & s = s_2^A \\ 1, & s = s_2^B \end{cases} \qquad \pi^2(s) = \begin{cases} 2, & s = s_1 \\ 1, & s = s_2^A \\ 2, & s = s_2^B \end{cases}$$

where $\pi^i : S \mapsto A$ denotes the action selected by player i at each state. Here, $V^{\pi^1,\pi^2}(s_1) = 1+0 = 1$, whereas $V^{\pi^2,\pi^1}(s_1) = -1 + 1 = 0$. The game value is asymmetric, violating MSG, despite every local payoff matrix being skew-symmetric.

This counter-example clearly demonstrates that the MSG assumption induces severe structural constraints. When players swap policies (switching π^1 and π^2), the resulting trajectory of state visitations may change completely. Yet, the MSG condition demands that the total value of the game remains exactly symmetric (negated) despite this potentially drastically different path through the state space. This implies that the payoffs and transition dynamics must be tightly coupled to ensure such invariance holds across all possible Markov policies.

H.2 MSG but not HSG

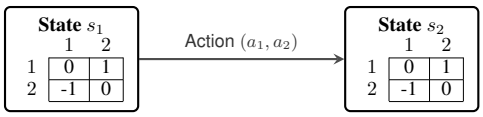


Figure 2: Example: MSG but not HSG game.

Consider the game above with deterministic transitions $P(s_2 \mid s_1, \cdot, \cdot) = 1$. Let us examine the value under specific history-dependent strategies:

$$\pi^1(z) = \begin{cases} 1, & z = (s_1) \\ 1, & z = (s_1, 1, 2, s_2) \\ 1, & z = (s_1, 2, 1, s_2) \end{cases}; \qquad \pi^2(z) = \begin{cases} 2, & z = (s_1) \\ 1, & z = (s_1, 1, 2, s_2) \\ 2, & z = (s_1, 2, 1, s_2) \end{cases},$$

where $\pi^i : Z \mapsto A$ denotes the action selected by player i for each history.

Under these strategies, the trajectory is $(s_1, 1, 2) \to s_2$. The value is: $V^{\pi^1,\pi^2}(s_1) = u(s_1, 1, 2) + u(s_2, 1, 1) = 1 + 0 = 1$. However, if we swap roles, the trajectory becomes $(s_1, 2, 1) \to s_2$, and the opponent responds differently to this history. The value becomes: $V^{\pi^2,\pi^1}(s_1) = u(s_1, 2, 1) + u(s_2, 2, 1) = -1 - 1 = -2$. Since $1 \neq -(-2)$, the game violates HSG. This example demonstrates that HSG requires symmetry to hold even when strategies condition their future behavior on the specific history of past actions. This restriction forces the continuation value to be independent of future policies after a disagreement has occurred.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification:

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.