Thompson Sampling for Markov Games with Piecewise Stationary Opponent Policies

Anthony DiGiovanni[1] **Ambuj Tewari**[1]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Abstract

Reinforcement learning problems with multiple agents pose the challenge of efficiently adapting to nonstationary dynamics arising from other agents' strategic behavior. Although several algorithms exist for these problems with promising empirical results, regret analysis and efficient use of other-agent models in general-sum games is very limited. We propose an algorithm (TSMG) for general-sum Markov games against agents that switch between several stationary policies, combining change detection with Thompson sampling to learn parametric models of these policies. Under standard assumptions for parametric Markov decision process (MDP) learning, the expected regret of TSMG in the worst case over policy parameters and switch schedules is near-optimal in time and number of switches, up to logarithmic factors. Our experiments on simulated games show that TSMG can outperform standard Thompson sampling and a version of Thompson sampling with a static reset schedule, despite the violation of an assumption that the MDPs induced by the other player are ergodic.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) seeks to apply optimal sequential decision making methods to problems involving other intelligent actors. Markov (or stochastic) games provide a formalism that is commonly used to study MARL. In particular, learning algorithms for *general-sum* games are appropriate for interactions in which players are neither fully cooperative nor competitive, and thus have the potential to avoid costly conflicts and achieve mutually beneficial outcomes. These algorithms have applications in contexts such as autonomous vehicle navigation [Tang, 2019], where each plan to reach a destination must account for the trajectories of other drivers, and delegated bargaining [Tossou et al., 2020]. Most theoretical progress in this field is limited to zero-sum games [Wei et al., 2017, Bai and Jin, 2020, Fan et al., 2020], where the other player (hereafter "player 2") is purely adversarial, or to games where players share the same reward function [Asghari et al., 2020]. In noncooperative general-sum games, predicting the other player's policy is more difficult, since its reward-maximizing policy at a given time depends on the alignment of this agent's objectives with the learner's, and beliefs about the learner. Analysis of finite-time performance (regret) in multi-agent settings, beyond asymptotic convergence to equilibria, faces the challenge that player 2's learning induces a nonstationary environment. While weak results can be guaranteed by optimizing one's policy with respect to a pessimistic hypothesis that player 2 is adversarial, general-sum games require more sophisticated models of player 2 for near-optimal rewards.

One important model for this problem is the piecewise stationary Markov decision process (MDP) with parametric structure. In certain games, it is useful to hypothesize that player 2's policy is from a parametric family, constructed with prior knowledge about this agent's goals and algorithm, with parameters that change throughout the game as player 2 learns [He et al., 2016, Everett and Roberts, 2017]. For example, in games where certain expert strategies such as those explored in tournaments [Axelrod, 1984] are common, one can learn more efficiently than through nonparametric estimates by modeling player 2 as switching between these strategies. Policy gradient agents can also be modeled parametrically [Foerster et al., 2018, Letcher et al., 2019], and recent single-agent literature has considered linear models of transition dynamics [Ayoub et al., 2020]. Section 5.3 gives an example of a model for a class of strategies used in repeated games. In these settings, it is crucial for a learning agent to quickly adapt to the other agent's changes, to avoid exploitative attempts to "teach" mistaken beliefs about future behavior. It is easy to provide an example Markov

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

game where failure to adapt results in linear regret (e.g., see Appendix A).

Here, we consider the problem of a parametric piecewise stationary Markov decision process, motivated by the application of this model to Markov games against a player 2 that switches policies. Works such as Hernandez-Leal et al. [2014, 2016b] and Radanovic et al. [2019] have considered similar settings, but from an empirical perspective. Our result is analogous to the near-optimal regret bound of MASTER [Wei and Luo, 2021] for the nonparametric setting, with the potential for greater efficiency due to the use of parametric structure. We take a model-based approach, combining a change detection procedure with a Thompson sampling reinforcement learning (RL) algorithm for parametric MDPs.

2 RELATED WORK

Model-based RL with parametric transition models is well-studied, for time-homogeneous MDPs. The Thompson sampling algorithm TSMDP [Gopalan and Mannor, 2015] for a non-changing MDP achieves worst-case regret $\tilde{\mathcal{O}}(T^{1/2})$ with probability $1-\delta$ (ignoring problem-dependent factors independent of T, and suppressing logarithmic factors in $\tilde{\mathcal{O}}$ notation). With weaker assumptions, DS-PSRL [Theocharous et al., 2018] has Bayes regret $\tilde{\mathcal{O}}(CH(C'T)^{1/2})$ in an analogous setting, where C and C' are constants governing smoothness of the MDP dynamics and concentration of the posterior, respectively, and H is a bound on the span of the differential value function. Using upper confidence exploration, in episodic MDPs of episode length H with model family Eluder dimension d, UCRL-VTR [Ayoub et al., 2020] has worst-case regret $\tilde{\mathcal{O}}(H\min\{d,T\}+(dT)^{1/2}+HT^{1/2})$ with probability $1-\delta$. Our algorithm aims for similar worst-case regret scaling in T, and no multiplicative dependence on S or A as in the result for TSMDP, in the parametric piecewise stationary setting.

The piecewise stationary MDP problem was considered briefly in the analysis of UCRL2 [Jaksch et al., 2010]. To our knowledge, that work provides the only regret bound for discretely changing MDPs, specifically $\tilde{\mathcal{O}}(DSA^{1/2}\ell^{1/3}T^{2/3})$ for $\ell-1$ changes. This algorithm follows a static schedule of times to reset UCRL2. By contrast, our algorithm achieves a lower rate in T (thus a lower overall rate when $\ell=o(T)$) by actively detecting changes. A recent work, using multiple instances of UCRL2 managed by a meta-algorithm called MASTER, achieves the same rate in ℓ and T as ours for finite-diameter MDPs [Wei and Luo, 2021]. However, their algorithm is for the nonparametric setting and therefore has dependence on S and A in the dominant term; since the nonparametric model is a special case of the parametric model with a multinomial distribution, ours is more general for ergodic piecewise stationary MDPs. Additionally, MASTER requires optimistic value estimates from its base algorithm, thus it is not clear that Thompson sampling base algorithms can be used. Thompson sampling tends to work better than optimism-based RL algorithms in practice [Osband and Van Roy, 2017]. Finally, since no numerical experiments are provided in Wei and Luo [2021], it is unknown how practically successful and computationally efficient MASTER is.

In piecewise stationary bandits, M-UCB [Cao et al., 2019] achieves near-optimal regret $\tilde{\mathcal{O}}((\ell KT)^{1/2})$ with change detection similar to ours, where K is the number of arms. Though ℓ is not assumed known, M-UCB's regret bound requires a hyperparameter that depends on ℓ. Banerjee et al. [2017] propose an algorithm for changing MDPs, using a two-threshold strategy that temporarily switches to an information-maximizing policy when a small change is detected. In their setting, however, the full transition models before and after each change are known. EXPDRBIAS [Radanovic et al., 2019] achieves *external* regret (with respect to the best stationary policy rather than a sequence) $\mathcal{O}(T^{\max\{1-3\alpha/7,1/4\}})$ in episodic Markov games, for $\alpha>0$ such that player 2's largest policy change between episodes scales as $\mathcal{O}(T^{-\alpha})$. While this is close to our setting and applies to a diverse class of other players, this result is for a weaker notion of regret than ours, and requires full observation of player 2's policy at the end of each episode.

Recent literature has attempted to solve general-sum Markov games using switching player 2 models. FAL-SG [Elidrisi et al., 2014] uses an experts algorithm to adapt to changing policies, with each expert selecting from an updated finite set of hypotheses about player 2's strategy. While this approach has the important property of never doing much worse on average than the maximin value of the game, it does not guarantee low regret against nonstationary players. The Bayesian algorithm OLSI [Hernandez-Leal and Kaisers, 2017] updates beliefs over a set of hypothesized models in games against switching agents, and DriftER [Hernandez-Leal et al., 2016b] uses explicit change detection with resets. While empirically successful in both classic games and more realistic tasks, these algorithms again lack regret analysis.

3 PROBLEM DESCRIPTION AND ALGORITHM

3.1 PRELIMINARIES

We consider the model of a 2-player general-sum continuing Markov game, a tuple $(\mathcal{S},\mathcal{A}^{(1)},\mathcal{A}^{(2)},r^{(1)},\mathcal{T})$, as a parametric piecewise stationary MDP. Although our results can easily be extended to general piecewise stationary MDPs, due to our motivating examples, we focus on the case where the dynamics of the piecewise stationary MDP can be factored into an unchanging part and a changing

part, where the latter has the structure of a policy. This game is defined by the state space $\mathcal{S}$, action spaces of the respective players $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$, player 1's reward function $r^{(1)}$, and transition dynamics $\mathcal{T}$. Player 2's rewards $r^{(2)}$ will not appear in our analysis, but knowing $r^{(2)}$ may be useful for designing models of player 2's policy. Assume $S := |\mathcal{S}|$ and $A := |\mathcal{A}^{(1)}| = |\mathcal{A}^{(2)}|$ are finite, and that for all $s \in \mathcal{S}, a^{(1)} \in \mathcal{A}^{(1)}, a^{(2)} \in \mathcal{A}^{(2)}$, the $r^{(1)}(s, a^{(1)}, a^{(2)}) \in [0, 1]$ and $\mathcal{T}(\cdot|s, a^{(1)}, a^{(2)})$ are known.[1] Let Θ be a space of parameters, where $\theta \in \Theta$ parameterizes player 2's policy $\pi^{(2)}_\theta$. We suppose that throughout the Markov game, player 2 follows a sequence of ℓ fixed policies parameterized by $\{\theta^*_1, ..., \theta^*_\ell\}$. Both this sequence and the times $\{\nu_1, ..., \nu_{\ell-1}\}$ at which player 2 switches policies are unknown to player 1, our learning agent. The number of time steps T and ℓ are also unknown, although we will see that knowing these values permits a choice of a hyperparameter for our algorithm necessary to achieve the regret bound. Therefore player 1 faces a piecewise stationary MDP, where each stationary phase m has induced transition dynamics $P_{\theta_m}(\cdot|s, a^{(1)}) = \sum_{a^{(2)} \in \mathcal{A}^{(2)}} \mathcal{T}(\cdot|s, a^{(1)}, a^{(2)}) \pi^{(2)}_{\theta_m}(a^{(2)}|s)$.

Let $(S_i, A^{(1)}_i, A^{(2)}_i, R_i)$ be the state, actions taken, and reward $R_i = r^{(1)}(S_i, A^{(1)}_i, A^{(2)}_i)$ to player 1 at time step i. Partially adopting the notation in Gopalan and Mannor [2015], let $\mathcal{C}$ be the space of stationary, deterministic policies over $\mathcal{S} \times \mathcal{A}^{(1)}$. For each $\pi^{(1)} \in \mathcal{C}$ and time horizon $t \in \mathbb{N}$, define $H_{t,\theta,\pi^{(1)}} : \mathcal{S} \to \mathbb{R}$ for an MDP induced by θ as $H_{t,\theta,\pi^{(1)}}(s) := \mathbb{E}_{\theta,\pi^{(1)}}\left[\sum_{i=0}^{t} R_i | S_0 = s\right]$. Let $\pi^{\text{OPT}(1)}(\theta) := \arg\max_{\pi^{(1)} \in \mathcal{C}} \lim_{t\to\infty} \frac{1}{t} H_{t,\theta,\pi^{(1)}}(s_0)$, i.e., the policy with optimal long-term average reward given θ, and $\mu^{\text{OPT}}(\theta) := \max_{\pi^{(1)} \in \mathcal{C}} \lim_{t\to\infty} \frac{1}{t} H_{t,\theta,\pi^{(1)}}(s_0)$. We choose an arbitrary initial state s_0, because we assume each MDP is ergodic (Assumption 2 below), and in such MDPs the optimal average reward does not depend on the initial state [Puterman, 1994]. Define $\nu_0 := 0$ and $\nu_\ell := T$. Then, for fixed sequences $\{\theta^*_1, ..., \theta^*_\ell\}$ and $\{\nu_1, ..., \nu_{\ell-1}\}$ we define regret (for the sequence of rewards produced by execution of a given algorithm) in this problem as follows:

$$\mathcal{R}(T) := \sum_{m=1}^{\ell} (\nu_m - \nu_{m-1}) \mu^{\text{OPT}}(\theta^*_m) - \sum_{t=0}^{T-1} R_t.$$

That is, regret is the gap in rewards between player 1 and an oracle that knows $\{\theta^*_1, ..., \theta^*_\ell\}$ and $\{\nu_1, ..., \nu_{\ell-1}\}$. Player 1's goal is to play a sequence of policies minimizing expected regret, in the worst case over sequences of θ^*_m and ν_m satisfying our assumptions, that is, minimize $\max_{\{\theta^*_1,...,\theta^*_\ell\},\{\nu_1,...,\nu_{\ell-1}\}} \mathbb{E}(\mathcal{R}(T))$.

[1]The assumption of known rewards is for expositional convenience, as in related work; it is well-known that MDP regret bounds increase only by a constant factor when this assumption is relaxed [Bartlett and Tewari, 2009, Gopalan and Mannor, 2015, Agrawal and Jia, 2017].

This is stricter than regret with respect to a stationary policy, but since player 2's sequence of policies is fixed in this definition, it is weaker than comparing to the sequence of policies that a strategic player 2 *would have used* in response to "optimal" play [Crandall, 2014]. We use this definition to leave the analysis fully general, without an explicit model of how player 2 chooses to switch policies. Though we use Thompson sampling for this objective, we defer analysis of Bayes regret to future work. We aim to nearly match the lower bound of $\Omega((\ell T)^{1/2})$ for piecewise stationary MDPs, which follows from applying the stationary MDP bound of $\Omega((DSAT)^{1/2})$ [Jaksch et al., 2010] to phases of length $\lceil T/\ell \rceil$.

3.2 THOMPSON SAMPLING FOR MARKOV GAMES WITH CHANGE DETECTION (TSMG)

Given this model of player 2's policies, player 1 begins the game with a prior p_Θ over the parameter defining player 2's first policy. We use Thompson sampling [Thompson, 1933] to balance exploration and exploitation. Over a sequence of epochs, player 1 uses the history of player 2's actions to update p_Θ to a posterior distribution $P_{\Theta|H}$ with Bayes' rule. At the start of an epoch, player 1 samples an estimate $\hat{\theta} \sim P_{\Theta|H}$, and follows the optimal policy for an MDP induced by $\hat{\theta}$ until the next epoch. After at least L time steps, an epoch concludes at the first return time to a fixed state s_0, which is positive recurrent under every stationary optimal policy with respect to the prior (see Assumption 2). The minimum length guarantees sufficiently many samples from the same player 1 policy, for use in change detection.

With the exception of this extra criterion of minimum epoch length, we use TSMDP [Gopalan and Mannor, 2015] as the base RL algorithm. To account for player 2's switches, we add change detection, such that with high probability player 1 forgets previous (irrelevant) data if and only if player 2 recently changed policies. Sequential analysis offers several candidate algorithms. Though the generalized likelihood ratio (GLR) method [Banerjee et al., 2017] suits parametric models, to our knowledge the finite-sample results are insufficient to prove the false positive and negative rates we require. Given the form of each $\pi^{(2)}$, a procedure based on the beliefs computed in subsequent epochs would exploit the parametric structure, potentially with asymptotic no-regret guarantees. However, similarly, we lack concentration inequalities for general parametric families that would enable proof of performance. While our experiments evaluate a parametric change detection procedure as well, for theoretical tractability we use simple nonparametric test statistics (Algorithm 1) in the algorithm we analyze. Specifically, this procedure compares the empirical frequencies of player 2's actions conditioned on states between epochs. Given sufficiently sparse switches and large "distance" between subsequent player 2 policies, elaborated below, a properly

tuned threshold for this procedure provides low false positive and false negative rates that guarantee our regret bound.

Algorithm 1 CD

Require: Histories $H = (S_1, A_1^{(2)}, ..., S_{N_1}, A_{N_1}^{(2)})$ and $H' = (S'_1, A_1'^{(2)}, ..., S'_{N_2}, A_{N_2}'^{(2)})$, threshold b

Compute $N_{s,a} \leftarrow \sum_{t=1}^{N_1} \mathbb{I}[S_t = s, A_t^{(2)} = a]$,
$N'_{s,a} \leftarrow \sum_{t=1}^{N_2} \mathbb{I}[S'_t = s, A_t'^{(2)} = a]$,
$N_s \leftarrow \sum_{t=1}^{N_1} \mathbb{I}[S_t = s]$, $N'_s \leftarrow \sum_{t=1}^{N_2} \mathbb{I}[S_t = s]$,
$\hat{M}_{s,a} \leftarrow \frac{N_{s,a}}{N_s} \mathbb{I}[N_s > 0] + \frac{1}{A} \mathbb{I}[N_s = 0]$,
$\hat{M}'_{s,a} \leftarrow \frac{N'_{s,a}}{N'_s} \mathbb{I}[N'_s > 0] + \frac{1}{A} \mathbb{I}[N'_s = 0]$.
if $|||\hat{M} - \hat{M}'||| > b$ **then**
    Return `true`
Return `false`

Algorithm 2 TSMG

Init s_0, $P_{\Theta|H} \leftarrow p_\Theta$, $e(t) \leftarrow 0$, $u \leftarrow 0$
for $t = 0, 1, ..., T-1$ **do**
    if $s_t = s_0$ and $u \in \{0\} \cup [L, \infty)$ **then**
        $e(t) \leftarrow e(t) + 1$
        $u \leftarrow 0$
        if CD$(H(e(t)-2), H(e(t)-1), b) =$ `true` and $e(t) \geq 3$ **then**
            $P_{\Theta|H} \leftarrow p_\Theta$
        Sample $\hat{\theta} \sim P_{\Theta|H}$
        $\pi^{(1)} \leftarrow \pi^{\text{OPT}(1)}(\hat{\theta})$
    $a_t^{(1)} \leftarrow \pi^{(1)}(s_t)$
    $P_{\Theta|H} \leftarrow$ `Bayes`$(P_{\Theta|H}, s_t, a_t^{(2)})$
    $u \leftarrow u + 1$

We briefly explain Algorithm 2, TSMG. The index u tracks whether an epoch has exceeded its minimum length (line 3). Let $e(t)$ be the epoch index at time t and $H(k)$ be the data history from epoch k. Change detection is not conducted until epoch 3 (line 6), since this requires 2 epochs of data. TSMG proceeds similarly to TSMDP, except that if a change is detected, the posterior is reset to the prior (line 7). TSMG can use any algorithm for computing $\pi^{\text{OPT}(1)}$ given a sampled estimate (line 9); we assume $\mathcal{S}$, $\mathcal{A}^{(1)}$, and $\mathcal{A}^{(2)}$ are sufficiently small that this computation is feasible. Access to an optimal policy oracle is standard in literature on Thompson sampling in MDPs [Gopalan and Mannor, 2015, Ouyang et al., 2017, Theocharous et al., 2018]. Since each optimal policy is deterministic, we denote by $\pi^{(1)}(s)$ the action taken in state s (line 10). `Bayes` denotes the update of the posterior by Bayes' rule, with likelihood given by the model $\pi^{(2)}$ (line 11). In practice, only approximate posterior sampling and optimal policy computation appear necessary for good results, as we see in the experiments.

4 ANALYSIS

4.1 NOTATION

Let $\overline{\tau}_{m,\pi^{(1)}} := \mathbb{E}_{\theta_m^*,\pi^{(1)}}(\min\{t \geq 1 | S_t = s_0\} | S_0 = s_0)$ be the expected recurrence time to s_0 under $(\theta_m^*, \pi^{(1)})$, and $\overline{\tau}^* := \min_{m=1,...,\ell}\{\overline{\tau}_{m,\pi^{\text{OPT}(1)}(\theta_m^*)}\}$. As our algorithm builds on TSMDP, whose analysis relies on these recurrence times, our analysis also involves these problem-dependent quantities. Let $\hat{\nu}_m$ be the mth time at which TSMG infers a change, that is, resets the posterior to the prior. The events $F_m := [e(\hat{\nu}_m) > e(\nu_m)]$ and $D_m := [e(\hat{\nu}_m) \leq e(\nu_m) + 2]$ are, respectively, the case that no false positives and no more than one false negative occur for the mth change, for $m = 1, ..., \ell$. Define $B_m := \cap_{i=1}^{m-1} F_i D_i$, that is, the occurrence of all of these "good" events up to and excluding phase m.

Suppose that B_{m+1} holds, meaning CD works near-optimally for the first m policy changes. Then the regret incurred by player 1 after $\hat{\nu}_m$ is bounded by that of a hypothetical game *starting at time* ν_m with $\ell - m$ player 2 phases remaining, since at time $\hat{\nu}_m$ player 1's posterior is reset to the prior, and $\hat{\nu}_m \geq \nu_m$. Define $\tilde{\mathbb{E}}_m$ as the expectation for this remaining game starting at s_0 with a non-updated prior, conditional on B_{m+1}, where $\tilde{\mathbb{E}}_0 := \mathbb{E}$. Let $\tilde{T}_m := T - \nu_m$ and $\tilde{\nu}_{m+1} := \nu_{m+1} - \nu_m$ be, respectively, the effective time horizon and "first" switch time of the remaining game. Define t_k as the start time of epoch k, and the norm $|||M||| := \max_i \sum_j |M_{i,j}|$. Finally, we define $\epsilon = \min_{m=2,...,\ell} |||M_{(m)} - M_{(m-1)}|||$, where each $M_{(m)}$ is a matrix whose (s, a) entry is $\pi^{(2)}_{\theta_m^*}(a|s)$.

4.2 ASSUMPTIONS

The success of TSMG relies on two key conditions: first, that TSMDP will work well during periods in which player 2's policy is stationary, and second, that these periods are sufficiently long that player 1 has time to learn the new policy model after detecting a change. Assumption 1 formalizes these conditions. Though this appears restrictive, in certain general-sum games it is reasonable to model player 2 with a slowly switching policy. For example, when player 2 is an advanced opponent, with enough knowledge about player 1 that it does not need to explore or learn much, player 2 is incentivized to switch policies only when player 1 rejects an attempt to "teach" a policy benefiting player 2 [Crandall, 2020]. Hence, it does not help player 2 to switch so frequently that player 1 can't learn. In a single-agent nonstationary setting, Cao et al. [2019] make an analogous assumption. Non-theoretical papers on learning against switching opponents with similar assumptions include Hernandez-Leal et al. [2016a] and Everett and Roberts [2017].

Assumption 1. *Let T_m be the minimum number of time*

steps required such that Theorem 2 of Gopalan and Mannor [2015] holds, for the MDP induced by θ^*_m. Define the number of epochs of phase m (rounding up) by $E_m := e(\nu_m) - e(\nu_{m-1})$. Then for all $m = 1, ..., \ell$, for the value L used to set the epoch stopping times, we have that $E_m \geq \max\{\frac{T_m}{L}, 2\}$.

Because TSMG relies on the strong performance of a subroutine resembling TSMDP, conditional on no false positives by CD, we require each condition necessary for the regret bound of TSMDP in a *stationary* MDP. In particular, Assumption 2 and the technical conditions deferred to Appendix B must hold for each MDP induced by $\{\theta^*_1, ..., \theta^*_\ell\}$, for fixed s_0. The positive recurrence assumption is relatively strong in a multi-agent problem where states represent past joint actions (see Section 5). However, we will see from empirical results that this does not appear strictly necessary for the desired regret rate. This condition holds for irreducible games, where almost surely every state is visited infinitely often for any pair of policies [Neyman and Sorin, 2003].

Assumption 2. *The start state s_0 is positive recurrent for the true MDP induced by θ^*_m, under each $\pi^{\text{OPT}(1)}(\theta) \in \mathcal{C}$ for θ in the support of p_Θ.*

4.3 MAIN RESULTS

Algorithm 2 incurs low regret by executing efficient RL while player 2 is stationary, and quickly discarding irrelevant data when player 2 changes strategies. Lemma 1 shows this first property: given no false positives between a reset time and player 2's next policy change, the regret during this segment matches that of TSMDP in a non-changing MDP.

Lemma 1. *For each $m = 1, ..., \ell$, we have:*

$$\tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{\nu}_m)\mathbb{I}[F_m]) \leq \mathcal{O}\left(\sqrt{\frac{\tilde{\nu}_m}{\overline{\tau}^*}\log\left(2\tilde{\nu}_m \log(\tilde{\nu}_m)\right)}\right).$$

Proof. Given F_m and B_m (implicit in the notation $\tilde{\mathbb{E}}_{m-1}$), the posterior is never reset during the interval $[\nu_{m-1}, \nu_m]$ in which player 2 follows a stationary policy. Thus in this interval, the induced MDP is time-homogeneous, and player 1 follows an algorithm equivalent to TSMDP. By Assumption 1, $\nu_m - \nu_{m-1}$ is sufficiently large that Theorem 2 of Gopalan and Mannor [2015] holds with $\overline{\tau}_{m,\pi^{\text{OPT}(1)}(\theta^*_m)}$ as the recurrence time and $\tilde{\nu}_m = \nu_m - \nu_{m-1}$ as the time horizon. That is, the regret bound for TSMDP fails to hold with probability no greater than δ. By the definition of $\overline{\tau}^*$, choosing $\delta = \frac{1}{\tilde{\nu}_m}$:

$$\begin{aligned}\tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{\nu}_m)\mathbb{I}[F_m]) &\leq \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{\nu}_m)|F_m) \\ &\leq \delta\tilde{\nu}_m + \mathcal{O}\left(\sqrt{\frac{\tilde{\nu}_m}{\overline{\tau}^*}\log\left(\frac{2\log(\tilde{\nu}_m)}{\delta}\right)}\right) \\ &= \mathcal{O}\left(\sqrt{\frac{\tilde{\nu}_m}{\overline{\tau}^*}\log\left(2\tilde{\nu}_m\log(\tilde{\nu}_m)\right)}\right).\end{aligned}$$

□

Extending the analysis of Markov chain estimation from Wolfer and Kontorovich [2019], Lemma 2 provides a concentration inequality for nonparametric estimation of player 2's policy, which will help establish the accuracy of CD. The proof is in Appendix C.

Lemma 2. *For any epoch k in which player 2 follows a stationary policy, denote the policy matrix for that epoch M_k and empirical estimate $\hat{M}_k$ computed from n samples of the MDP induced by M_k. Note the distinction from $M_{(m)}$ for each **phase**. Let $\mathcal{M}_k$ be the Markov chain transition matrix induced by M_k and player 1's policy $\pi^{(1)}$ in epoch k. This chain has a stationary distribution p_k because the state space is finite and at least one state is positive recurrent. Let γ_k be the pseudo-spectral gap [Paulin, 2015] of $\mathcal{M}_k$, $p^*_k := \min_{s\in\mathcal{S}} p_k(s)$, $p^* := \inf_{k=1,2,...} p^*_k$, $\gamma := \inf_{k=1,2,...}\gamma_k$, and $p(s_0) := \inf_{k=1,2,...} p_k(s_0)$. Define $\eta(n,\gamma,p^*) := \frac{\gamma n^2 p^*}{8[8(n+\frac{1}{\gamma})(1-p^*)+10n]}$. Then, for any $x \in (0,2)$:*

$$\begin{aligned}&P(|||\hat{M}_k - M_k||| > x) \leq \\ &S\left[(A+1)\exp\left(-\frac{np^*x^2}{32A}\right) + \frac{\exp(-\eta(n,\gamma,p^*))}{\sqrt{p(s_0)}}\right].\end{aligned}$$

Given sufficiently long epochs, this lemma controls the rates of false positives and negatives, as formalized in Lemma 3. For false positives, the estimates of player 2's policy from subsequent epochs can only differ greatly from each other if at least one estimate differs greatly from the true $\pi^{(2)}_{\theta^*_m}$. A large sample size prevents this. Avoiding any false negatives is very challenging, since player 2 may switch toward the end of an epoch, making the data from that epoch and the preceding one almost indistinguishable. However, in this case, the data from that epoch will differ noticeably from those of the *next* epoch. Thus we avoid two false negatives in a row with high probability. The proof is in Appendix D.

Lemma 3. *Suppose that the threshold b is contained in the*

following interval:

$$\left[\sqrt{\frac{128A}{Lp^*}\log\left(\frac{A+1}{\frac{1}{2S(\ell T)^{1/2}(\lceil T/L\rceil-1)}-\frac{\exp(-\eta(L,\gamma,p^*))}{\sqrt{p(s_0)}}}\right)},\ \frac{\epsilon}{2}-\sqrt{\frac{32A}{Lp^*}\log\left(\frac{A+1}{\frac{1}{2S(\ell T)^{1/2}}-\frac{\exp(-\eta(L,\gamma,p^*))}{\sqrt{p(s_0)}}}\right)}\right].$$

Then $P(F_m^c|B_m) \leq \frac{1}{(\ell T)^{1/2}}$ *for* $m = 1, ..., \ell$, *and* $P(D_m^c F_m|B_m) \leq \frac{1}{(\ell T)^{1/2}}$ *for* $m = 1, ..., \ell - 1$.

Remark 4. *If ℓ is unknown, all instances of $(\ell T)^{1/2}$ in Lemma 3 can be replaced with T when computing an appropriate b. Theorem 5 will still hold in this case, as the leading terms in the regret bound are unaffected. Knowledge of ℓ provides slightly less stringent bounds on b.*

A practical limitation of Lemma 3 is that the constraints on the change detection threshold b depend on unknown statistics of the induced MDPs, the pseudo-spectral gap γ and stationary distribution p. These quantities can be estimated at the end of an epoch, however [Paulin, 2015, Combes and Touati, 2019], and we remark that the main competitor to our algorithm, MASTER, also requires knowledge of the maximum diameter of the MDPs (assuming ℓ is unknown) [Wei and Luo, 2021].

Lemmas 1 and 3 let us prove our main result, Theorem 5. The full expected regret consists of contributions from the ideal case, with no false positives and no more than one false negative for the first switch time, and the non-ideal case of either of these errors. We bound the probability of the non-ideal case such that total regret from this contribution scales as $\ell^{1/2}T^{1/2}$. In the ideal case, player 1 does as well as if playing a stationary MDP until the first switch, then potentially does poorly for no more than 2 epochs, and resets such that a "new" game begins with one less switch. This provides a recursion, adding analogous terms for a sequence of games of decreasing length, so that the dominant contributions to regret are those that player 1 would have had given a *known* schedule of switch times. Our regret bound generalizes Theorem 2 of Gopalan and Mannor [2015]. When $\ell = 1$, that is, no change occurs, player 2 is a parametric MDP and we have average regret $\tilde{\mathcal{O}}(T^{1/2})$ as expected.

Theorem 5. *Suppose there exists $b > 0$ satisfying Lemma 3. Then, running Algorithm 2 given Assumptions 1-5 satisfies:*

$$\max_{\substack{\{\theta_1^*,\ldots,\theta_\ell^*\}\\ \{\nu_1,\ldots,\nu_{\ell-1}\}}} \mathbb{E}(\mathcal{R}(T)) \leq \mathcal{O}(\ell + \sqrt{\ell T \log(T\log(T))}).$$

Proof. We follow a line of argument similar to Cao et al. [2019]. Fix $\{\theta_1^*, ..., \theta_\ell^*\}$ and $\{\nu_1, ..., \nu_{\ell-1}\}$. Consider each phase $m = 1, ..., \ell$. Given that epochs start at the positive recurrent state s_0, there exists a finite $C_\ell := \max_{m=1,\ldots,\ell} \tilde{\mathbb{E}}_{m-1}(t_{e(\nu_m)+2} - t_{e(\nu_m)}|F_m D_m)$, independent of T. Then, since $R_t \in [0,1]$ for all t and $(\hat{\nu}_m - \nu_{m-1}) - \tilde{\nu}_m = (\hat{\nu}_m - \nu_{m-1}) - (\nu_m - \nu_{m-1})$, for each $m = 1, ..., \ell - 1$:

$$\begin{aligned}
&\tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1}) - \mathcal{R}(\tilde{\nu}_m)|F_m D_m)\\
&\quad= \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1}) - \mathcal{R}(\hat{\nu}_m - \nu_{m-1})|F_m D_m)\\
&\quad+ \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\hat{\nu}_m - \nu_{m-1}) - \mathcal{R}(\tilde{\nu}_m)|F_m D_m)\\
&\quad\leq \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1}) - \mathcal{R}(\hat{\nu}_m - \nu_{m-1})|F_m D_m)\\
&\quad+ \tilde{\mathbb{E}}_{m-1}(\hat{\nu}_m - \nu_m|F_m D_m)\\
&\quad\leq \tilde{\mathbb{E}}_m(\mathcal{R}(\tilde{T}_m)) + \tilde{\mathbb{E}}_{m-1}(\hat{\nu}_m - \nu_m|F_m D_m) \qquad (3)\\
&\quad\leq \tilde{\mathbb{E}}_m(\mathcal{R}(\tilde{T}_m)) + \tilde{\mathbb{E}}_{m-1}(t_{e(\nu_m)+2} - t_{e(\nu_m)}|F_m D_m) \qquad (4)\\
&\quad\leq \tilde{\mathbb{E}}_m(\mathcal{R}(\tilde{T}_m)) + C_\ell.
\end{aligned}$$

Line 3 follows from the definition of $\tilde{\mathbb{E}}_m$. That is, given B_m and $F_m D_m$ we have B_{m+1}, and we bound the regret after $\hat{\nu}_m$ by the regret from a hypothetical game starting at time ν_m. Line 4 holds because by D_m, $e(\hat{\nu}_m) \leq e(\nu_m) + 2$. Next:

$$\begin{aligned}
&\tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1}) - \mathcal{R}(\tilde{\nu}_m))\\
&\quad\leq \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1}) - \mathcal{R}(\tilde{\nu}_m)|F_m D_m)\\
&\quad+ (\tilde{T}_{m-1} - \tilde{\nu}_m)P((F_m D_m)^c|B_m) \qquad (1)\\
&\quad\leq \tilde{\mathbb{E}}_m(\mathcal{R}(\tilde{T}_m)) + C_\ell + \tilde{T}_m P((F_m D_m)^c|B_m)\\
&\quad\leq \tilde{\mathbb{E}}_m(\mathcal{R}(\tilde{T}_m)) + C_\ell + \frac{2\tilde{T}_m}{(\ell T)^{1/2}}. \qquad (3)
\end{aligned}$$

In line 1 we again used boundedness of rewards, and bound $P(F_m D_m|B_m) \leq 1$. Line 3 follows from $P((F_m D_m)^c|B_m) = P(F_m^c|B_m) + P(D_m^c F_m|B_m)$ and Lemma 3. Further:

$$\begin{aligned}
&\tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1})) = \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1}) - \mathcal{R}(\tilde{\nu}_m))\\
&\quad+ \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{\nu}_m)\mathbb{I}[F_m^c]) + \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{\nu}_m)\mathbb{I}[F_m])\\
&\quad\leq \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{T}_{m-1}) - \mathcal{R}(\tilde{\nu}_m)) + \tilde{\nu}_m P(F_m^c|B_m)\\
&\quad+ \tilde{\mathbb{E}}_{m-1}(\mathcal{R}(\tilde{\nu}_m)\mathbb{I}[F_m])\\
&\quad\leq \tilde{\mathbb{E}}_m(\mathcal{R}(\tilde{T}_m)) + C_\ell + \frac{2\tilde{T}_m + \tilde{\nu}_m}{(\ell T)^{1/2}}\\
&\quad+ \mathcal{O}\left(\sqrt{\frac{\tilde{\nu}_m}{\overline{\tau}^*}\log\left(2\tilde{\nu}_m\log(\tilde{\nu}_m)\right)}\right). \qquad \text{(Lemma 1)}
\end{aligned}$$

Accounting for the regret after the last change time, and noting that $\tilde{T}_{\ell-1} = T - \nu_{\ell-1} = \nu_\ell - \nu_{\ell-1} = \tilde{\nu}_\ell$:

$$\begin{aligned}
&\tilde{\mathbb{E}}_{\ell-1}(\mathcal{R}(\tilde{T}_{\ell-1})) = \tilde{\mathbb{E}}_{\ell-1}(\mathcal{R}(\tilde{\nu}_\ell)\mathbb{I}[F_\ell^c]) + \tilde{\mathbb{E}}_{\ell-1}(\mathcal{R}(\tilde{\nu}_\ell)\mathbb{I}[F_\ell])\\
&\quad\leq \frac{\tilde{\nu}_\ell}{(\ell T)^{1/2}} + \mathcal{O}\left(\sqrt{\frac{\tilde{\nu}_\ell}{\overline{\tau}^*}\log\left(2\tilde{\nu}_\ell\log(\tilde{\nu}_\ell)\right)}\right).
\end{aligned}$$

Therefore, summing the contributions to regret from all ℓ phases, we have, by the telescoping sum $\sum_{m=1}^{\ell} \tilde{\nu}_m = T$:

$$\begin{aligned}
\mathbb{E}(\mathcal{R}(T)) &\leq (\ell-1)C_\ell - \frac{2\tilde{T}_\ell}{(\ell T)^{1/2}} + \sum_{m=1}^{\ell} \frac{2\tilde{T}_m + \tilde{\nu}_m}{(\ell T)^{1/2}} \\
&+ \sum_{m=1}^{\ell} \mathcal{O}\left(\sqrt{\frac{\tilde{\nu}_m}{\overline{\tau}^*} \log\left(\tilde{\nu}_m \log(\tilde{\nu}_m)\right)}\right) \\
&\leq (\ell-1)C_\ell + \sum_{m=1}^{\ell} \frac{2T}{(\ell T)^{1/2}} \\
&+ \sum_{m=1}^{\ell} \mathcal{O}\left(\sqrt{\frac{\tilde{\nu}_m}{\overline{\tau}^*} \log\left(\tilde{\nu}_m \log(\tilde{\nu}_m)\right)}\right) \\
&= \mathcal{O}\left(\ell + \sqrt{\ell T} + \sqrt{\ell \sum_{m=1}^{\ell} \frac{\tilde{\nu}_m}{\overline{\tau}^*} \log\left(\tilde{\nu}_m \log(\tilde{\nu}_m)\right)}\right) \\
&= \mathcal{O}(\ell + \sqrt{\ell T \log(T \log(T))}).
\end{aligned}$$

□

5 NUMERICAL EXPERIMENTS

Code for the experiments in this section is available on Github.[2] In each experiment, we use the following procedure. The parameter for each model family is in $\mathbb{R}^4$. Let $\mathbf{1}$ be the vector of ones. In experiments 1 and 2, the prior is Dirichlet with $\alpha = 0.5 \cdot \mathbf{1}$. The prior for experiment 3 is a vector of i.i.d. log-normals with $\mu = 0$ and $\sigma = 0.5$. We use the Metropolis algorithm for posterior sampling, with 3000, 2000, and 800 draws for experiments 1, 2, and 3, respectively (the empirical minimum needed for convergence). Each draw size n is preceded by $\frac{n}{4}$ burned samples, and we use the last draw as the sample for line 8 of Algorithm 2. In each game, there is no clear way to identify a state satisfying Assumption 2. We therefore test the sensitivity of TSMG to this condition, implementing it without checking in Line 3 for s_0.

We set $T = 10^5$, $L = 2500$, and $\ell = 6$. For each game, we make 10 schedules $\{\nu_1, ..., \nu_{\ell-1}\}$ by drawing uniformly at random from $\{1, 2, ..., 10^5 - 1\}$, rejecting draws until the spacing satisfies Assumption 1. For each schedule, we compute the average cumulative regret at each time step, over 30 games. The maximum of these averages provides an empirical bound on expected regret, since by Theorem 5 this bound should hold over all valid schedules for fixed ℓ. For empirical regret, optimal returns for each θ_m^* are computed as the average of an optimal policy's rewards over a rollout of sufficiently long time horizon (that is, T) that the initial state does not affect the average. We use value iteration to compute approximately optimal policies.

[2] https://github.com/digiovannia/tsmg

0.75, 0.75	0, 1
1, 0	0.25, 0.25

Prisoner's Dilemma

1, 0.5	0, 0
0, 0	0.5, 1

Bach-or-Stravinsky

1, 0.667	0.462, 0.5	0, 0.833
0.462, 0.5	0.615, 1	0, 0.833
0.769, 0	0.769, 0	0.308, 0.333

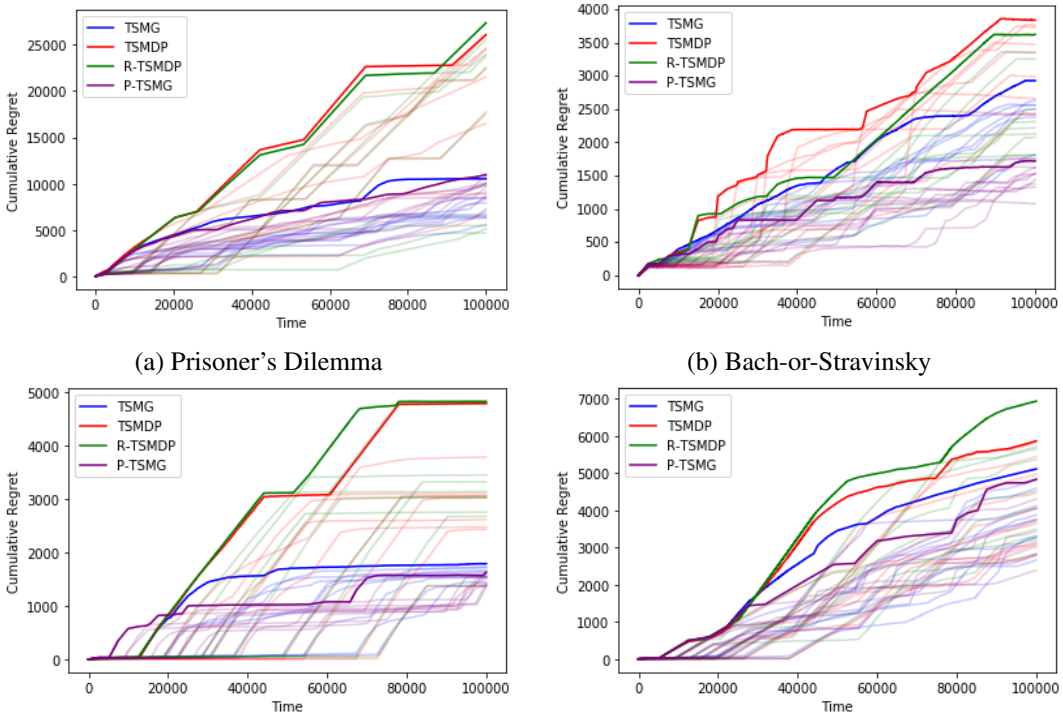
BOS+PD

Figure 1: Payoff matrices for Experiments 1 and 3.

The threshold for change detection is $b = 1.5$. This value is not chosen directly based on the bounds in Lemma 3, but in practice it achieves a reasonable rate of false positives and negatives. We also compare to three other baselines. First, Algorithm 2 is run without change detection (labeled TSMDP in all plots). Second, we modify this algorithm with no change detection to reset with the UCRL2 schedule defined in Theorem 6 of Jaksch et al. [2010] (R-TSMDP). That is, at each time $t = \lceil \frac{i^3}{\ell^2} \rceil$ for $i = 1, 2, ...$, the history $N_{s,a}$ and N_s are reset to 0. Third, we replace CD in TSMG with a parametric test (P-TSMG). A change is declared if $||\overline{\theta} - \overline{\theta}'||_2 > c$, where $\overline{\theta}, \overline{\theta}'$ are averages of the last $\frac{n}{2}$ posterior samples from the past two epochs. We set $c = 0.5$ for experiments 1 and 2, and $c = 4$ for experiment 3, to match the scale of the parameters.

5.1 EXPERIMENT 1: ITERATED PRISONER'S DILEMMA AND BACH-OR-STRAVINSKY

To compare with other algorithms for games with switching opponents, notably Hernandez-Leal et al. [2016a,b], we test TSMG in repeated versions of the Prisoner's Dilemma (PD) and Bach-or-Stravinsky (BOS). Figure 1 shows the payoff matrices for these games. In any state, $r^{(i)}(s, a^{(1)}, a^{(2)})$ is the ith value listed in the $(a^{(1)}, a^{(2)})$ cell of the matrix. States are given by the pair of each player's last two actions, i.e., the state $\{(i, j), (k, l)\}$ means that player 1's last two actions were i then j, and player 2's were k then l. Though the dynamics of a repeated game can depend heavily on the initial beliefs of the players, for simplicity in all our repeated games we let both players begin with a "memory" of action 1 in both previous turns, so the initial state is $\{(1, 1), (1, 1)\}$. Player 2 switches between weight vectors for linear combinations of a set of base policies for each game. (Compare with linear mixture model-based RL, as in Ayoub et al. [2020].) That is, for a set of policies $\{\pi_1^{(2)}, ..., \pi_d^{(2)}\}$ and $\theta \in \Delta^{d-1}$ in the probability simplex, we have $\pi_\theta^{(2)}(a|s) = \sum_{j=1}^{d} \theta_j \pi_j^{(2)}(a|s)$. The base policies, modeled after standard game-theoretic strategies in these games, are in Appendix E.1. For experiments 1 and 2, the values of θ_m^* are constructed as follows. Let e_i be the ith canonical vector. Player 2 cycles through $0.8 \cdot e_i + 0.05 \cdot \mathbf{1}$ for $i = 1, 2, 3, 4$ in order, wrapping around for $\ell > 4$. Thus in each phase, player 2 predominantly places weight on one



(a) Prisoner's Dilemma

(b) Bach-or-Stravinsky

(c) Grid

(d) BOS+PD

Figure 2: Results from Experiments 1-3. For each algorithm, the light curves are pointwise average regret from the different $\{\nu_m\}$ schedules, and the bolded curve is the pointwise maximum of the light curves.

of the base policies, but with some exploration of the others as well.

We abbreviate average regret in the worst case over schedules of ν as "regret" unless otherwise stated. In PD (Figure 2a), the regret of TSMG and P-TSMG increases roughly as $T^{1/2}$. TSMDP's and R-TSMDP's regret increases in a piecewise linear trend, with total regret far exceeding TSMG. For BOS (Figure 2b), however, TSMG's regret appears closer to linear and is outperformed by P-TSMG, although it is still lower than the baselines.

5.2 EXPERIMENT 2: GRID WORLD

Next, we evaluate TSMG in a grid world game designed in Hu and Wellman [2003] (Figure 3).[3] The state space is defined by the positions of the players. Cell 8 is a goal for both players, which can only be occupied by one player at a time, thus they face a potential conflict. The bottom corners are respective starting positions. Players return to the start upon reaching the goal, and are rewarded for reaching the goal but penalized for colliding with each other. The red lines in Figure 3 indicate barriers through which players only pass with probability 0.5. Appendix E.2 provides the formal rules of the grid and model for player 2's policies.

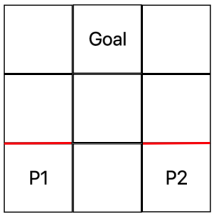


Figure 3: Grid world for Experiment 2.

TSMG and P-TSMG achieve $T^{1/2}$ regret scaling in this game (Figure 2c), contrasted with the piecewise linear scaling of the other baselines, with less than half the total regret of these algorithms.

5.3 EXPERIMENT 3: ADAPTIVE GODFATHER STRATEGY IN 3X3 GAME

Finally, we design a parametric model for player 2's policies in a repeated game, which generalizes the Godfather strategy [Littman and Stone, 2001]. This strategy proposes an

[3] We attempted to test TSMG in the other grid from Hu and Wellman [2003], however, player 1 was able to get optimal rewards across a wide range of inaccurate estimates $\hat{\theta}$. Thus this task was not sufficiently challenging to be appropriate for our experiments, and false positives would be much more costly than false negatives.

outcome in the game, which in general maximizes the user's reward but may be modified for "fairness," and punishes the other player for deviating from this proposal. Behavior is governed by a vector $\theta = (\theta_B, \theta_P, \theta_E, \theta_F) \in \mathbb{R}^4_+$. Increasing θ_B increases player 2's tendency toward the Bully policy (see Appendix E.1). Increasing θ_P promotes punishment of actions by player 1 that result in lower reward for player 2, or, if θ_E is high, lower product of both players' rewards (quantifying how *egalitarian* the outcome is). Increasing θ_F makes player 2 more forgiving, i.e., less likely to punish two turns in a row. The full model is given in Appendix E.3.

Player 2 cycles through $\{\theta^*_m\} = \{(1, 0, 0, 0), (1, 10, 0, 0), (1, 10, 1, 0), (1, 10, 1, 5), (1, 5, 5, 0), (0, 10, 0, 0)\}$. Player 2 consistently affords weight to the Bully strategy, but tests a variety of punishment policies before also testing pure punishment. The game used for this experiment is a combination of PD and an asymmetric form of BOS (BOS+PD, Figure 1). Both players prefer the top-left and middle cells to bottom-right, yet the unique stage game Nash equilibrium is (3, 3), similar to PD. Further, player 1 prefers one of these two cells while player 2 prefers the other, as in BOS.

In BOS+PD, the regret scaling of TSMG and P-TSMG appears roughly piecewise square-root. Though this trend is not starkly distinguishable from those of TSMDP and R-TSMDP, the latter two algorithms incur higher regret. P-TSMG slightly improves upon TSMG as well.

From these experiments, we have seen that even when Assumption 2 is violated, TSMG can outperform algorithms that either do not account for dynamics changes, or only passively reset without inferring such changes from data. The empirical scaling of TSMG's regret over time is consistent with our analysis in all games except BOS. Although Appendix A showed that in general the regret of TSMDP in our setting may scale as $\Omega(T)$, in these games TSMDP sometimes shows a similar scaling to TSMG, or outperforms R-TSMDP, which we would expect to have regret $\tilde{\mathcal{O}}(T^{2/3})$. Regardless, TSMG is a strict improvement. In Appendix F, we evaluate TSMG in self-play, where strong performance is not theoretically guaranteed.

6 DISCUSSION

We designed the TSMG algorithm for Markov games against a player 2 with parametric piecewise stationary policies. TSMG uses Thompson sampling for planning during periods of stationarity and, with change detection, adapts to switches of player 2's strategy through appropriately timed forgetting of old data. Under conditions on the feasibility of learning the Markov chains induced by both players' policies, we proved a competitive regret bound for TSMG. Although nonparametric change detection has a high sample complexity, and one of these conditions is relatively restrictive, our experiments demonstrate the robustness of TSMG to violation of this condition. TSMG outperforms two competitor algorithms, and a parametric modification of TSMG provides a greater improvement despite the absence of a regret guarantee.

Remaining open challenges for MARL theory include combination of the adaptability of TSMG with provable self-play compatibility, low regret against adaptive opponents (as in Crandall [2014]), and robustness across a variety of model classes for opponents.

Author Contributions

A. D. wrote the paper and code for numerical experiments. A. T. proposed the problem and helped edit the paper.

Acknowledgements

A. D. acknowledges the support of a grant from the Center on Long-Term Risk Fund. A. T. acknowledges the support of NSF via grant IIS-2007055.

References

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

Seyed Mohammad Asghari, Yi Ouyang, and Ashutosh Nayyar. Regret bounds for decentralized learning in cooperative multi-agent dynamical systems. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 2020.

Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F. Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120, pages 666–686, 2020.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 551–560, 2020.

Taposh Banerjee, Miao Liu, and Jonathan P. How. Quickest change detection approach to optimal control in markov decision processes with model changes. *American Control Conference*, pages 399–405, 2017.

Peter L. Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, page 35–42, 2009.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pages 418–427, 2019.

Richard Combes and Mikael Touati. Computationally efficient estimation of the spectral gap of a markov chain. *SIGMETRICS Perform. Eval. Rev.*, 47(1):98–100, 2019.

Jacob W. Crandall. Towards minimizing disappointment in repeated games. *Journal of Artificial Intelligence Research*, 49:111–142, 2014.

Jacob W. Crandall. When autonomous agents model other agents: An appeal for altered judgment coupled with mouths, ears, and a little more tape. *Artificial Intelligence*, 280, 2020.

Mohamed Elidrisi, Nicholas Johnson, Maria Gini, and Jacob Crandall. Fast adaptive learning in repeated stochastic games by game abstraction. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, page 1141–1148, 2014.

Richard Everett and Stephen Roberts. Learning against non-stationary agents with opponent modelling and deep reinforcement learning. *NIPS Workshop on Learning in the Presence of Strategic Behavior*, 2017.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120, pages 486–489, 2020.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, page 122–130, 2018.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 861–898, 2015.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1804–1813, 2016.

Pablo Hernandez-Leal and Michael Kaisers. Learning against sequential opponents in repeated stochastic games. *The 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making*, 2017.

Pablo Hernandez-Leal, Enrique Munoz de Cote, and L. Enrique Sucar. Using a priori information for fast learning against non-stationary opponents. In *Proceedings of the 16th Ibero-American Conference on Artificial Intelligence*, page 536–547, 2014.

Pablo Hernandez-Leal, Matthew E. Taylor, Benjamin Rosman, L. Enrique Sucar, and Enrique Munoz de Cote. Identifying and tracking switching, non-stationary opponents: A bayesian approach. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence Multiagent Interaction without Prior Coordination*, 2016a.

Pablo Hernandez-Leal, Yusen Zhan, Matthew E. Taylor, L. Enrique Sucar, and Enrique Munoz de Cote. Efficiently detecting switches against non-stationary opponents. *Autonomous Agents and Multi-Agent Systems*, 31:767–789, 2016b.

Junling Hu and Michael P. Wellman. Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069, 2003.

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

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*, 2019.

Michael Littman and Peter Stone. Implicit negotiation in repeated games. In *Proceedings of The Eighth International Workshop on Agent Theories, Architectures, and Languages*, page 393–404, 2001.

Olivier Marchal and Julyan Arbel. On the sub-gaussianity of the beta and dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017.

Abraham Neyman and S. Sorin, editors. *Stochastic Games and Applications*. Springer Netherlands, 2003.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2701–2710, 2017.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems 30*, page 1333–1342, 2017.

Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

Rob Powers and Yoav Shoham. New criteria and a new algorithm for learning in multi-agent systems. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, page 1089–1096, 2004.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

Goran Radanovic, Rati Devidze, David C. Parkes, and Adish Singla. Learning to collaborate in markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5261–5270, 2019.

Yichuan Charlie Tang. Towards learning multi-agent negotiations via self-play. *ICCV Workshop*, 2019.

Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. Scalar posterior sampling with applications. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, page 7696–7704, 2018.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 24:285–294, 1933.

Aristide C.Y. Tossou, Christos Dimitrakakis, Jaroslaw Rzepecki, and Katja Hofmann. A novel individually rational objective in multi-agent multi-armed bandits: Algorithms and regret bounds. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, page 1395–1403, 2020.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach, 2021. URL `https://arxiv.org/abs/2102.05406`.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, page 4994–5004, 2017.

Geoffrey Wolfer and Aryeh Kontorovich. Minimax learning of ergodic markov chains. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98, pages 904–930, 2019.

Thompson Sampling for Markov Games with Piecewise Stationary Opponent Policies (Supplementary Material)

Anthony DiGiovanni[1] **Ambuj Tewari**[1]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, USA

A EXAMPLE MARKOV GAME WHERE TSMDP HAS LINEAR REGRET

We show that there exists a game where TSMDP, without change detection, incurs linear regret. In this example, the data from a sequence of player 2 policies in which θ^* switches between two extremes produce a posterior heavily concentrated at the average of those extremes. In half of the phases after an initial period, the estimates $\hat{\theta}$ bias player 1 away from the optimal policy with high probability.

Consider an iterated game of Bach-or-Stravinsky (see Section 5.1). Given parameter $\theta^*_m \in [0,1]$, player 2's policy is to play action 2 with probability θ^*_m, independent of the state. That is, the actions $X_i := \mathbb{I}[A_i^{(2)} = 2]$ taken by player 2 for $i = \tau_{m-1}, ..., \tau_m - 1$ are independent draws from Bernoulli(θ^*_m). Let player 1's prior be Beta(α, β). Then at each time t, the posterior used by TSMDP (which treats θ^* as fixed) is Beta$(\alpha + \sum_{i=0}^{t-1} X_i, \beta + t - \sum_{i=0}^{t-1} X_i)$. Given that $\pi^{(2)}_{\theta^*_m}$ and $r^{(1)}$ are independent of state, so is player 1's optimal action at a given time. Thus given an estimate $\hat{\theta}$, and supposing WLOG ties are broken in favor of action 1, player 1 plays action 1 if and only if $\frac{1}{2}\hat{\theta} \leq 1 - \hat{\theta}$, that is, $\hat{\theta} \leq \frac{2}{3}$.

Suppose that for all $m = 1, ..., \ell$, we have $\nu_m - \nu_{m-1} = K$ for some K satisfying Assumption 1. For odd m, let $\theta^*_m = 1$ and for even m, let $\theta^*_m = 0$. Then for odd m, $\mu^{\text{OPT}}(\theta^*_m) = \frac{1}{2}$ and for even m, $\mu^{\text{OPT}}(\theta^*_m) = 1$. Let ℓ be even. Hence:

$$\begin{aligned}
\mathcal{R}(T) &= \sum_{m=1}^{\ell} (\nu_m - \nu_{m-1}) \mu^{\text{OPT}}(\theta^*_m) - \sum_{t=0}^{T-1} R_t \\
&= \frac{3K\ell}{4} - \sum_{m=1}^{\ell} \sum_{t=\nu_{m-1}}^{\nu_m - 1} R_t \\
&= \frac{3K\ell}{4} - \sum_{n=1}^{\ell/2} \left(\sum_{t=\nu_{2n-2}}^{\nu_{2n-1}-1} R_t + \sum_{t'=\nu_{2n-1}}^{\nu_{2n}-1} R_{t'} \right) \\
&= \frac{3K\ell}{4} - \sum_{n=1}^{\ell/2} \left(\frac{1}{2} \sum_{t=\nu_{2n-2}}^{\nu_{2n-1}-1} \mathbb{I}[\hat{\theta}_t > 2/3] + \sum_{t'=\nu_{2n-1}}^{\nu_{2n}-1} \mathbb{I}[\hat{\theta}_t \leq 2/3] \right).
\end{aligned}$$

Since player 2 always plays action 2 in an odd phase, and action 1 in even, then for t in an odd phase, the posterior is Beta$\left(\alpha + t - \frac{\lfloor t/K \rfloor K}{2}, \beta + \frac{\lfloor t/K \rfloor K}{2}\right)$, and for even, Beta$\left(\alpha + \frac{(\lfloor t/K \rfloor + 1)K}{2}, \beta + t - \frac{(\lfloor t/K \rfloor + 1)K}{2}\right)$. Now, let $I_x(a,b)$ denote the CDF of a Beta random variable with parameters a and b. By Theorem 1 of Marchal and Arbel [2017] (sub-Gaussianity of the Beta distribution with $\sigma^2_{opt}(a,b) < \frac{1}{4(a+b+1)}$), for $x > \frac{a}{a+b}$ we have:

$$1 - I_x(a,b) \leq \exp\left(-2(a+b+1) \left(x - \frac{a}{a+b} \right)^2 \right).$$

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Further, for $a = \alpha + t - \frac{\lfloor t/K \rfloor K}{2}$ and $b = \beta + \frac{\lfloor t/K \rfloor K}{2}$, we have $\frac{a}{a+b} \leq \frac{7}{12} < \frac{2}{3}$ whenever $t \geq 5\alpha - 7\beta + 6K$, because:

$$\begin{aligned}
t &\geq 5\alpha - 7\beta + 6K, \\
\frac{7t}{12} &\geq \frac{5}{12}\alpha - \frac{7}{12}\beta + \frac{t+K}{2}, \\
\frac{7}{12}(\alpha + \beta + t) &\geq \alpha + t - \frac{t-K}{2}, \\
\frac{7}{12} &\geq \frac{\alpha + t - \frac{t-K}{2}}{a+b} \\
&\geq \frac{\alpha + t - \frac{\lfloor t/K \rfloor K}{2}}{a+b} \\
&= \frac{a}{a+b}.
\end{aligned}$$

Let $T_{\alpha,\beta,K} := \max\{72\log(T) - \alpha - \beta - 1, 5\alpha - 7\beta + 6K\}$. Thus whenever $t \geq T_{\alpha,\beta,K}$,

$$\exp\left(-2(\alpha + \beta + t + 1)\left(\frac{2}{3} - \frac{\alpha + t - \frac{\lfloor t/K \rfloor K}{2}}{\alpha + \beta + t}\right)^2\right) \leq \frac{1}{T},$$

because:

$$\begin{aligned}
t &\geq -72\log(1/T) - \alpha - \beta - 1, \\
\log(1/T) &\geq -\frac{2}{144}(\alpha + \beta + t + 1), \\
\frac{1}{T} &\geq \exp\left(-2(\alpha + \beta + t + 1)\left(\frac{2}{3} - \frac{7}{12}\right)^2\right) \\
&\geq \exp\left(-2(\alpha + \beta + t + 1)\left(\frac{2}{3} - \frac{\alpha + t - \frac{\lfloor t/K \rfloor K}{2}}{\alpha + \beta + t}\right)^2\right).
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathbb{E}(\mathcal{R}(T)) &= \frac{3K\ell}{4} - \sum_{n=1}^{\ell/2}\left(\frac{1}{2}\sum_{t=\nu_{2n-2}}^{\nu_{2n-1}-1} P(\hat{\theta}_t > 2/3) + \sum_{t'=\nu_{2n-1}}^{\nu_{2n}-1} P(\hat{\theta}_t \leq 2/3)\right) \\
&\geq \frac{3K\ell}{4} - \sum_{n=1}^{\ell/2}\left[\frac{1}{2}\sum_{t=\nu_{2n-2}}^{\nu_{2n-1}-1} P(\hat{\theta}_t > 2/3) + K\right] \\
&\geq \frac{K\ell}{4} - \frac{1}{2}\left[\sum_{t=0}^{\lceil T_{\alpha,\beta,K}\rceil - 1} P(\hat{\theta}_t > 2/3) + \sum_{t=\lceil T_{\alpha,\beta,K}\rceil}^{T} \exp\left(-2(\alpha + \beta + t + 1)\left(\frac{2}{3} - \frac{\alpha + t - \frac{\lfloor t/K \rfloor K}{2}}{\alpha + \beta + t}\right)^2\right)\right] \\
&\geq \frac{T}{4} - \frac{1}{2}\left(T_{\alpha,\beta,K} + 1 + \sum_{t=\lceil T_{\alpha,\beta,K}\rceil}^{T} \frac{1}{T}\right) \\
&\geq \frac{T}{4} - \frac{1}{2}\left(T_{\alpha,\beta,K} + 2\right) \\
&\geq \min\left\{\frac{T}{4} - 36\log(T) + \frac{\alpha + \beta - 1}{2}, \frac{T}{4} - \frac{5\alpha - 7\beta + 6K + 2}{2}\right\} \\
&= \Omega(T).
\end{aligned}$$

B ASSUMPTIONS OF TSMDP

Time indices are modified from the source [Gopalan and Mannor, 2015] to be appropriate for the phases, or subgames, given by the sequence of player 2 policies. We refer the reader to Gopalan and Mannor [2015] for intuitive descriptions of the

assumptions. Further, we omit the source's Assumption 3 (uniqueness of the optimal policy) as the authors note it is not technically necessary, only useful for exposition.

Assumption 3. *There exists a constant $\Gamma < \infty$ such that: For all $\theta \in \Theta, (s, s', a^{(1)}) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}^{(1)}$, if $p_\Theta(\theta) > 0$ and $P_{\theta_m^*}(s'|s, a^{(1)}) > 0$, then $\left| \log \frac{P_{\theta_m^*}(s'|s,a^{(1)})}{P_\theta(s'|s,a^{(1)})} \right| \leq \Gamma$.*

Assumption 4. *For a given epoch j, let $\hat{\theta}_j$ be the parameter sampled by TSMG for that epoch; $p_{m,s_1,s_2}^{(\pi^{(1)})}$ be the stationary joint probability of s_1 immediately followed by s_2 under the Markov chain induced by $\pi^{(1)}$ and $\pi_{\theta_m^*}^{(2)}$; and $P_{\Theta|H_{t_k}}$ be the posterior at time t_k, that is, conditioned on all data between the last reset time and t_k. Define $N_{\pi^{(1)}}(k) := \sum_{j=e(\nu_{m-1})}^{k} \mathbb{I}[\pi^{\text{OPT}(1)}(\hat{\theta}_j) = \pi^{(1)}]$ and $J_{s_1,s_2}(k, \pi^{(1)}) := \sum_{t=\nu_{m-1}}^{\nu_m} \mathbb{I}[\pi^{\text{OPT}(1)}(\hat{\theta}_{e(t)}) = \pi^{(1)} \cap (S_t, S_{t+1}) = (s_1, s_2) \cap N_{\pi^{(1)}}(e(t)) \leq k]$. Suppose $e_1, e_2 \geq 0$. Then there exists $p^*(e_1, e_2) > 0$ such that, for any epoch index $k = \sum_{\pi^{(1)} \in \mathcal{C}} k_{\pi^{(1)}}$ in which the following holds for all $s_1, s_2 \in \mathcal{S}$, $k_{\pi^{(1)}} \geq 1$, and $\pi^{(1)} \in \mathcal{C}$:*

$$\left| \frac{J_{s_1,s_2}(k_{\pi^{(1)}}, \pi^{(1)})}{k_{\pi^{(1)}}} - \overline{\tau}_{m,\pi^{(1)}} p_{m,s_1,s_2}^{(\pi^{(1)})} \right| \leq \sqrt{\frac{e_1 \log(e_2 \log(k_{\pi^{(1)}}))}{k_{\pi^{(1)}}}},$$

we have $P_{\Theta|H_{t_k}}(\{\theta \in \Theta \mid \pi^{\text{OPT}(1)}(\theta) = \pi^{\text{OPT}(1)}(\theta_m^*)\}) \geq p^*(e_1, e_2)$.

Assumption 5. *Define the marginal Kullback-Leibler divergence for θ under policy $\pi^{(1)}$:*

$$D_{\pi^{(1)}}(\theta_m^* || \theta) := \sum_{s_1 \in \mathcal{S}} \left(\sum_{s_2' \in \mathcal{S}} p_{m,s_1,s_2'}^{(\pi^{(1)})} \right) KL(P_{\theta_m^*}(\cdot|s_1, \pi^{(1)}(s_1)) || P_\theta(\cdot|s_1, \pi^{(1)}(s_1))).$$

***(A)** There exist $a_1 > 0$, $a_2 \geq 0$ such that for all choices of nonnegative integers $\{k_{\pi^{(1)}}, \pi^{(1)} \in \mathcal{C}\}$ and $k = \sum_{\pi^{(1)} \in \mathcal{C}} k_{\pi^{(1)}}$:*

$$p_\Theta \left(\left\{ \theta \in \Theta \;\middle|\; \sum_{\pi^{(1)} \in \mathcal{C}} k_{\pi^{(1)}} \overline{\tau}_{m,\pi^{(1)}} D_{\pi^{(1)}}(\theta_m^* || \theta) \leq 1 \right\} \right) \geq a_1 k^{-a_2}.$$

***(B)** There exist $a_3 > 0$, $a_4 > 0$ such that for all choices of nonnegative integers $\{k_{\pi^{(1)}}, \pi^{(1)} \in \mathcal{C}\}$ and $k = \sum_{\pi^{(1)} \in \mathcal{C}} k_{\pi^{(1)}}$ such that $k_{\pi^{(1)\text{OPT}}(\theta_m^*)} \geq k - 3\log^2(k)$:*

$$p_\Theta \left(\left\{ \theta \in \Theta \;\middle|\; \sum_{\pi^{(1)} \in \mathcal{C}} k_{\pi^{(1)}} \overline{\tau}_{m,\pi^{(1)}} D_{\pi^{(1)}}(\theta_m^* || \theta) \leq 1 \right\} \right) \geq a_3 k^{-a_4}.$$

C PROOF OF LEMMA 2

A direct application of Theorem 1 from Wolfer and Kontorovich [2019] to the Markov chain induced by both players' policies would not be sufficient, since we would be unable to separate the effect of player 1's change in policy between epochs from that of a possible change in player 2's policy. Although one could condition on player 1's actions as well, estimating a transition dynamics tensor, the upper bound would scale as S^2 rather than S. Thus the structure of the changing MDP as a sequence of changing *policies*, with observable actions by player 2, permits more efficient estimation than in a general piecewise stationary MDP.

Lemma 2. *For any epoch k in which player 2 follows a stationary policy, denote the policy matrix for that epoch M_k and empirical estimate $\hat{M}_k$ computed from n samples of the MDP induced by M_k. Note the distinction from $M_{(m)}$ for each **phase**. Let $\mathcal{M}_k$ be the Markov chain transition matrix induced by M_k and player 1's policy $\pi^{(1)}$ in epoch k. This chain has a stationary distribution p_k because the state space is finite and at least one state is positive recurrent. Let γ_k be the pseudo-spectral gap [Paulin, 2015] of $\mathcal{M}_k$, $p_k^* := \min_{s \in \mathcal{S}} p_k(s)$, $p^* := \inf_{k=1,2,\ldots} p_k^*$, $\gamma := \inf_{k=1,2,\ldots} \gamma_k$, and $p(s_0) := \inf_{k=1,2,\ldots} p_k(s_0)$. Define $\eta(n, \gamma, p^*) := \frac{\gamma n^2 p^*}{8[8(n+\frac{1}{\gamma})(1-p^*)+10n]}$. Then, for any $x \in (0, 2)$:*

$$P(|||\hat{M}_k - M_k||| > x) \leq S \left[(A+1) \exp\left(-\frac{np^* x^2}{32A} \right) + \frac{\exp(-\eta(n, \gamma, p^*))}{\sqrt{p(s_0)}} \right].$$

Proof. For consistency with Wolfer and Kontorovich [2019], in this proof we index states by i. For an initial distribution μ_k define $\Pi_{\mu_k,k} := \sum_{i\in\mathcal{S}} \frac{\mu_k(i)^2}{p_k(i)}$. Since all epochs start at s_0, we have $\Pi_{\mu_k,k} = \frac{1}{p_k(s_0)}$. The proof of Theorem 1 in Wolfer and Kontorovich [2019] goes through, replacing the resulting state indexing of columns with action indexing, and setting $n_i = \frac{np_k(i)}{2}$ for each state i. Specifically, if we define $Y_{t,a}(k,i) = \frac{1}{\sqrt{2}}\mathbb{I}[S_t = i](\mathbb{I}[A_t^{(2)} = a] - M_k(i,a))$, their proof establishes that $\{\mathbf{Y}_t(k,i)\}$ is a martingale difference sequence. By the definitions of p_k^* and p^*, we have $p^* \le p_k^* \le p_k(i)$ for all states i and epochs k. Thus Corollary 1.3 from Tropp [2011] gives:

$$\begin{aligned} P(||\hat{M}_k(i,\cdot) - M_k(i,\cdot)||_1 > x \cap N_i \in [n_i, 3n_i]) &\le (A+1)\exp\left(-\frac{x^2 n_i^2}{8A(3n_i + xn_i/(3\sqrt{2A}))}\right) \\ &\le (A+1)\exp\left(-\frac{np^*x^2}{32A}\right). \end{aligned}$$

Applying the proof of Lemma 6 in Wolfer and Kontorovich [2019] to our case, which uses Theorem 3.4 and Proposition 3.10 from Paulin [2015], we have:

$$\begin{aligned} P(N_i \notin [n_i, 3n_i]) &\le \sqrt{\Pi_{\mu_k,k}}\exp\left(-\frac{\gamma_k(\frac{np_k(i)}{2})^2}{2[8(n+\frac{1}{\gamma_k})p_k(i)(1-p_k(i)) + 10np_k(i)]}\right) \\ &= \frac{1}{\sqrt{p_k(s_0)}}\exp\left(-\frac{\gamma_k n^2 p_k(i)}{8[8(n+\frac{1}{\gamma_k})(1-p_k(i)) + 10n]}\right) \\ &\le \frac{1}{\sqrt{p(s_0)}}\exp\left(-\frac{\gamma n^2 p^*}{8[8(n+\frac{1}{\gamma})(1-p^*) + 10n]}\right). \end{aligned}$$

Therefore, by the union bound:

$$\begin{aligned} P(|||\hat{M}_k - M_k||| > x) &= P\left(\max_{i\in S}\sum_{a\in\mathcal{A}^{(2)}} |\hat{M}_k(i,a) - M_k(i,a)| > x\right) \\ &= P\left(\bigcup_{i\in S}\left[\sum_{a\in\mathcal{A}^{(2)}} |\hat{M}_k(i,a) - M_k(i,a)| > x\right] \cap [\forall i, N_i \in [n_i, 3n_i]]\right) \\ &+ P\left(\bigcup_{i\in S}\left[\sum_{a\in\mathcal{A}^{(2)}} |\hat{M}_k(i,a) - M_k(i,a)| > x\right] \cap [\exists i, N_i \notin [n_i, 3n_i]]\right) \\ &\le \sum_{i\in S} P\left(||\hat{M}_k(i,\cdot) - M_k(i,\cdot)||_1 > x \cap N_i \in [n_i, 3n_i]\right) + P\left(\exists i, N_i \notin [n_i, 3n_i]\right) \\ &\le S\left[(A+1)\exp\left(-\frac{np^*x^2}{32A}\right) + \frac{1}{\sqrt{p(s_0)}}\exp(-\eta(n,\gamma,p^*))\right]. \end{aligned}$$

□

D PROOF OF LEMMA 3

Lemma 3. *Suppose that the threshold b is contained in the following interval:*

$$\left[\sqrt{\frac{128A}{Lp^*}\log\left(\frac{A+1}{\frac{1}{2S(\ell T)^{1/2}(\lceil T/L\rceil - 1)} - \frac{\exp(-\eta(L,\gamma,p^*))}{\sqrt{p(s_0)}}}\right)},\right.$$
$$\left.\frac{\epsilon}{2} - \sqrt{\frac{32A}{Lp^*}\log\left(\frac{A+1}{\frac{1}{2S(\ell T)^{1/2}} - \frac{\exp(-\eta(L,\gamma,p^*))}{\sqrt{p(s_0)}}}\right)}\right].$$

Then $P(F_m^c|B_m) \le \frac{1}{(\ell T)^{1/2}}$ for $m = 1, ..., \ell$, and $P(D_m^c F_m|B_m) \le \frac{1}{(\ell T)^{1/2}}$ for $m = 1, ..., \ell - 1$.

Proof. Given B_m, for $m > 1$ we have $e(\hat{\nu}_m) \geq e(\hat{\nu}_{m-1}) + 1 \geq e(\nu_{m-1}) + 2$. Since tests are not conducted until 2 epochs' data are collected, $e(\hat{\nu}_1) \geq 3 = e(\nu_0) + 2$. Thus:

$$\begin{aligned}
P(F_m^c|B_m) &= P(e(\hat{\nu}_m) \leq e(\nu_m)|B_m) \\
&= P\left(\bigcup_{j=e(\hat{\nu}_{m-1})+1}^{e(\nu_m)} [e(\hat{\nu}_m) = j] \middle| B_m\right) \\
&\leq P\left(\bigcup_{j=e(\nu_{m-1})+2}^{e(\nu_m)} [e(\hat{\nu}_m) = j] \middle| B_m\right) \\
&= \sum_{j=e(\nu_{m-1})+2}^{e(\nu_m)} P\left(|||\hat{M}_{j-2} - \hat{M}_{j-1}||| > b \cap \left(\bigcap_{k=e(\nu_{m-1})+2}^{j-1} |||\hat{M}_{k-2} - \hat{M}_{k-1}||| \leq b\right) \middle| B_m\right). \qquad (4)
\end{aligned}$$

Line 4 follows because, if we know that $e(\hat{\nu}_m) \geq e(\nu_{m-1}) + 2$ by B_m, the mth change is declared at some epoch $j \in \{e(\nu_{m-1}) + 2, ..., e(\nu_m)\}$ if and only if Algorithm 1 returned `False` in the previous epochs in that range, and returns `True` at that epoch. The probabilities of these mutually exclusive events are summed. Next:

$$\begin{aligned}
P(F_m^c|B_m) &\leq \sum_{j=e(\nu_{m-1})+2}^{e(\nu_m)} P(|||\hat{M}_{j-2} - \hat{M}_{j-1}||| > b|B_m) \\
&\leq \sum_{j=e(\nu_{m-1})+2}^{e(\nu_m)} [P(|||\hat{M}_{j-2} - M_{(m)}||| > b/2|B_m) + P(|||\hat{M}_{j-1} - M_{(m)}||| > b/2|B_m)] \\
&\leq 2(e(\nu_m) - e(\nu_{m-1}) - 1) \max_{k=e(\nu_{m-1}),...,e(\nu_m)-1} P(|||\hat{M}_k - M_{(m)}||| > b/2|B_m) \\
&\leq 2(\lceil T/L \rceil - 1) \max_{k=e(\nu_{m-1}),...,e(\nu_m)-1} P(|||\hat{M}_k - M_{(m)}||| > b/2|B_m). \qquad (4)
\end{aligned}$$

Because L is the minimum epoch length and thus the epoch index cannot exceed $\lceil T/L \rceil$, line 4 follows. Since each epoch between $e(\nu_{m-1})$ and $e(\nu_m) - 1$ features the same policy by player 2, there is one constant policy matrix $M_{(m)}$ throughout this sequence of epochs. The probability is almost of the form in Lemma 2, except conditioned on B_m. However, the random matrix $\hat{M}_k$ depends on B_m only through the properties of the induced Markov chain $\mathcal{M}_k$ and initial distribution $\mu_k(s) = \mathbb{I}[s = s_0]$. This dependence can be absorbed into the quantities $p(s_0)$, p^*, and γ used in the bound. Then by Lemma 2, since the length of each epoch is at least L and by hypothesis b is sufficiently large, we have:

$$\begin{aligned}
P(F_m^c|B_m) &\leq 2(\lceil T/L \rceil - 1)S \left[(A+1)\exp\left(-\frac{Lp^*b^2}{128A}\right) + \frac{1}{\sqrt{p(s_0)}}\exp(-\eta(L,\gamma,p^*))\right] \\
&\leq 2(\lceil T/L \rceil - 1)S \cdot \frac{1}{2S(\ell T)^{1/2}(\lceil T/L \rceil - 1)} \\
&= \frac{1}{(\ell T)^{1/2}}.
\end{aligned}$$

Next, by the definition of ϵ and the triangle inequality:

$$\begin{aligned}
|||\hat{M}_{e(\nu_m)-1} - M_{(m)}||| + |||\hat{M}_{e(\nu_m)} - M_{(m+1)}||| &\geq |||M_{(m)} - M_{(m+1)}||| - |||\hat{M}_{e(\nu_m)-1} - \hat{M}_{e(\nu_m)}||| \\
&> \epsilon - |||\hat{M}_{e(\nu_m)-1} - \hat{M}_{e(\nu_m)}|||.
\end{aligned}$$

And, also by the triangle inequality:

$$|||\hat{M}_{e(\nu_m)+1} - M_{(m+1)}||| \geq |||\hat{M}_{e(\nu_m)} - M_{(m+1)}||| - |||\hat{M}_{e(\nu_m)} - \hat{M}_{e(\nu_m)+1}|||.$$

Then, using the inequalities above in lines 5 and 6, respectively, since B_m implies $e(\hat{\nu}_m) \geq e(\nu_{m-1}) + 2$:

$$
\begin{aligned}
P(D_m^c F_m | B_m) &= P(e(\hat{\nu}_m) > e(\nu_m) + 2 \cap e(\hat{\nu}_m) > e(\nu_m) | B_m) \\
&\leq P\left(\bigcap_{j=e(\nu_{m-1})+2}^{e(\nu_m)+2} [e(\hat{\nu}_m) \neq j] \middle| B_m \right) \\
&\leq P(|||\hat{M}_{e(\nu_m)-1} - \hat{M}_{e(\nu_m)}||| \leq b \cap |||\hat{M}_{e(\nu_m)} - \hat{M}_{e(\nu_m)+1}||| \leq b | B_m) \\
&\leq P(|||\hat{M}_{e(\nu_m)-1} - M_{(m)}||| + |||\hat{M}_{e(\nu_m)} - M_{(m+1)}||| > \epsilon - b \cap |||\hat{M}_{e(\nu_m)} - \hat{M}_{e(\nu_m)+1}||| \leq b | B_m) \qquad (5) \\
&\leq P(|||\hat{M}_{e(\nu_m)-1} - M_{(m)}||| + |||\hat{M}_{e(\nu_m)+1} - M_{(m+1)}||| > \epsilon - 2b | B_m) \qquad (6) \\
&\leq P\left(|||\hat{M}_{e(\nu_m)-1} - M_{(m)}||| > \frac{\epsilon - 2b}{2} \middle| B_m \right) + P\left(|||\hat{M}_{e(\nu_m)+1} - M_{(m+1)}||| > \frac{\epsilon - 2b}{2} \middle| B_m \right).
\end{aligned}
$$

Given Assumption 1, if a policy change occurs in epoch $e(\nu_m)$, then no changes occur in $e(\nu_m) - 1$ and $e(\nu_m) + 1$, so $\hat{M}_{e(\nu_m)-1}$ and $\hat{M}_{e(\nu_m)+1}$ are computed from data purely produced by policies $M_{(m)}$ and $M_{(m+1)}$, respectively. Then we can apply Lemma 2, using the hypothesis that b is sufficiently small:

$$
\begin{aligned}
P(D_m^c F_m | B_m) &\leq 2S \left[(A+1) \exp\left(-\frac{Lp^*(\epsilon - 2b)^2}{128A} \right) + \frac{1}{\sqrt{p(s_0)}} \exp(-\eta(L, \gamma, p^*)) \right] \\
&\leq 2S \cdot \frac{1}{2S(\ell T)^{1/2}} \\
&= \frac{1}{(\ell T)^{1/2}}.
\end{aligned}
$$

□

E DETAILS ON GAMES FOR NUMERICAL EXPERIMENTS

E.1 PRISONER'S DILEMMA AND BACH-OR-STRAVINSKY

In Table 1, a given cell states the probability that player 2 takes action 1 in the row state if following the column pure policy. These base policies are as follows:

- Bully: plays the action that maximizes player 2's reward conditional on player 1 choosing the optimal action against that action
- Tit-for-Tat (TFT): plays the action player 1 played last turn
- Pavlov: plays 1 if both players played the same action last turn, otherwise 2
- Forgiving-TFT: plays 2 if and only if player 1 played 2 in both of the most recent turns
- Fair: plays both actions with equal probability
- Nash: plays the mixed-strategy Nash equilibrium of the matrix game
- Sequential (Seq): alternates playing its half of the two pure equilibria; if both players played the same action last turn, player 2 plays the opposite action, otherwise plays 2 (the action for the equilibrium that favors player 2)
- Forgiving-Seq: matches Seq, but if the players played different actions last turn yet the same action in the turn before, player 2 plays 2 (the equilibrium that favors player 1)

E.2 GRID WORLD

Actions are the cardinal directions. If a player's action is in the direction of an empty cell, the player moves deterministically to that cell, otherwise the player remains in place. These rules have 3 exceptions:

Table 1: State-Conditional Probability of Action 1 Under Base Policies for Experiment 1

State	Prisoner's Dilemma				Bach-or-Stravinsky			
	Bully	**TFT**	**Pavlov**	**Forgiving-TFT**	**Fair**	**Nash**	**Seq**	**Forgiving-Seq**
{(1,1), (1,1)}	0	1	1	1	$1/2$	$1/3$	1	1
{(1,1), (1,2)}	0	1	0	1	$1/2$	$1/3$	0	1
{(1,1), (2,1)}	0	1	1	1	$1/2$	$1/3$	1	1
{(1,1), (2,2)}	0	1	0	1	$1/2$	$1/3$	0	0
{(1,2), (1,1)}	0	0	0	1	$1/2$	$1/3$	0	1
{(1,2), (1,2)}	0	0	1	1	$1/2$	$1/3$	0	0
{(1,2), (2,1)}	0	0	0	1	$1/2$	$1/3$	0	0
{(1,2), (2,2)}	0	0	1	1	$1/2$	$1/3$	0	0
{(2,1), (1,1)}	0	1	1	1	$1/2$	$1/3$	1	1
{(2,1), (1,2)}	0	1	0	1	$1/2$	$1/3$	0	0
{(2,1), (2,1)}	0	1	1	1	$1/2$	$1/3$	1	1
{(2,1), (2,2)}	0	1	0	1	$1/2$	$1/3$	0	1
{(2,2), (1,1)}	0	0	0	0	$1/2$	$1/3$	0	0
{(2,2), (1,2)}	0	0	1	0	$1/2$	$1/3$	0	0
{(2,2), (2,1)}	0	0	0	0	$1/2$	$1/3$	0	1
{(2,2), (2,2)}	0	0	1	0	$1/2$	$1/3$	0	0

1. If a player i starts in the goal cell, any action results in the player returning to the start cell, unless the other player j moves to that cell, in which case i remains in the goal cell.

2. If players' actions result in a collision (aiming toward the same square or attempting to pass through each other), both remain in their starting cells.

3. Cells 1 and 3 have partial barriers north of them, such that if a player attempts to move north from one of these cells, the player passes only with probability 0.5, otherwise remains in place (independent of the other player's move).

To design a task for which an accurate world model is essential for player 1 to achieve minimal regret, we modify the rewards from Hu and Wellman [2003] to more heavily penalize collisions with player 2. Taking any action from the goal cell gives reward 1. Taking an action that results in player 1 and player 2 colliding gives reward 0. Any other state-action combination gives reward 0.7.

To construct policies for player 2, we use the following model. Player 1 knows that player 2's rewards from the goal cell and neutral actions are the same as player 1's, i.e., 1 and 0.7, respectively. However, player 1 is uncertain about the reward player 2 receives from a collision, which can take values in the set $\{0, 0.4, 0.6, 1\}$. On one extreme, player 2 has the same preferences as player 1. On the other, player 2 is adversarial, preferring a collision as much as reaching the goal cell.

Player 2 acts as if holding a distribution of weights θ over these possible reward values. Specifically, player 1 hypothesizes that for each reward value, player 2 knows the optimal policy with respect to the induced reward function and a fixed belief about player 1's policy, and plays a linear combination of these policies weighted by θ (just as in Experiment 1). Player 2's modeled belief is that player 1's policy is to take the shortest path to the goal cell (ignoring player 2's position). That is, under this belief, player 1 would always take the action in the direction of the goal cell when in the same row or column, and for any other cell, player 1 takes each of the two actions aimed toward the goal cell with probability 0.5.

This model can flexibly represent several patterns of player 2's behavior. Over the course of the game, represented by changes in θ, player 2 might undergo changes in intrinsic preferences for collisions, beliefs about the utility of "punishing" player 1 for thwarting player 2's goals, or desire to hide information from player 1 about the true reward function.

E.3 BOS+PD

For $s = (a^{(1)}, a^{(2)})$ and supposing both players' rewards are bounded within $[0, 1]$, we define the following features:

$$\phi_1(s,a) = r^{(2)}\left(\arg\max_{a'^{(1)}} r(a'^{(1)}, a), a\right), \qquad \text{(Bully)}$$

$$\phi_2(s,a) = 1 - \max_{a'^{(1)}} r^{(1)}(a'^{(1)}, a), \qquad \text{(Player 1's best possible reward from current action)}$$

$$\phi_3(s,a) = \max_{a'^{(1)}, a'^{(2)}} r^{(2)}(a'^{(1)}, a'^{(2)}) - \max_{a'^{(2)}} r^{(2)}(a^{(1)}, a'^{(2)}), \qquad \text{(Deviation from selfish proposal)}$$

$$\phi_4(s,a) = \max_{a'^{(1)}, a'^{(2)}} r^{(1)}(a'^{(1)}, a'^{(2)}) r^{(2)}(a'^{(1)}, a'^{(2)}) - \max_{a'^{(2)}} r^{(1)}(a^{(1)}, a'^{(2)}) r^{(2)}(a^{(1)}, a'^{(2)}),$$

(Deviation from egalitarian proposal)

$$\phi_5(s,a) = 1 - \max_{a'^{(1)}} r^{(1)}(a'^{(1)}, a^{(2)}). \qquad \text{(Player 1's best possible reward from past action)}$$

Then, in a given state s an action a is assigned the following score:

$$\begin{aligned}\rho(s,a|\theta) &= \theta_B \phi_1(s,a) + \theta_P[\phi_2(s,a)(\phi_3(s,a) + \theta_E \phi_4(s,a)) - \theta_F \phi_5(s,a)\phi_2(s,a)] \\ &= \theta_B \phi_1(s,a) + \theta_P \phi_2(s,a)\phi_3(s,a) + \theta_P \theta_E \phi_2(s,a)\phi_4(s,a) - \theta_P \theta_F \phi_5(s,a)\phi_2(s,a).\end{aligned}$$

Actions are selected according to a softmax distribution based on these scores:

$$\pi^{(2)}_\theta(a|s) = \frac{\exp(\rho(s,a|\theta))}{\sum_{a'} \exp(\rho(s,a'|\theta))}.$$

F SELF-PLAY EXPERIMENT

A standard desired criterion for a game-theoretic algorithm is "good" performance, for example convergence to a non-Pareto-dominated Nash equilibrium [Powers and Shoham, 2004], against a copy of itself, suggesting a stable incentive for multiple users to deploy this algorithm. Theorem 5 does not provide a guarantee for regret in self-play, since the policies produced by TSMG do not in general satisfy our assumptions. In particular, the learning agent may change policies as often as every epoch. Nonetheless, we examine our algorithm's self-play performance both due to this general motivation in game theory, and to test the robustness of TSMG to model misspecification.

For each game constructed in the previous experiments, both player 1 and player 2 model each other as described, and use the same prior, threshold b, and epoch length L. The indices of $a^{(1)}$ and $a^{(2)}$ are reversed in defining the features ϕ in Section E.3 when player 2 models player 1. We run each self-play game 20 times using two different priors. The "default" prior is the one used in the experiments of Section 5. We define "alternative" priors for each game: PD and BOS have $\alpha = (0.5, 0.5, 2, 0.5)$, grid has $\alpha = (2, 0.5, 0.5, 0.5)$, and BOS+PD has $\mu = (0, 2, 0, 0)$. In addition to regret, we also compare raw rewards gained by both players. This is because one weakness of the regret measure used in this work is that multiple equilibria can have the same (zero) regret despite drastically different average rewards.

We observe (Figure 4) that TSMG's performance in self-play is sometimes sensitive to the prior, and lower regret does not necessarily correspond to higher reward. Under both priors, the regret of TSMG in PD is sublinear. The alternative prior biases both players to expect each other to play the Pavlov strategy, and we find that Pavlov is an optimal policy against a Pavlov opponent. This allows the players to reach an equilibrium in which, absent false positives, they repeatedly gain the cooperation reward of 0.75, outperforming the copies following the default prior. However, cumulative regret under this alternative prior is actually higher.

In all other games, the regret is linear. For BOS, the prior biasing players toward believing each other will follow the Seq strategy does not improve performance. That is, players deploying TSMG who expect each other to place even moderate weight on the Seq strategy learn to fairly alternate between the two actions, achieving 0.75 on average each turn.

In the grid game, players who start expecting each other to place high weight on 0 reward from collision (that is, the reward corresponding to the largest entry of α) learn to avoid such collisions. Both players achieve higher reward consistently under this alternative prior.

Finally, the results of BOS+PD are similar to those of PD. The alternative prior places larger weight on the belief that the other player punishes deviations greatly, and this incentivizes both players to avoid the mutually costly bottom-right cell of the game matrix.

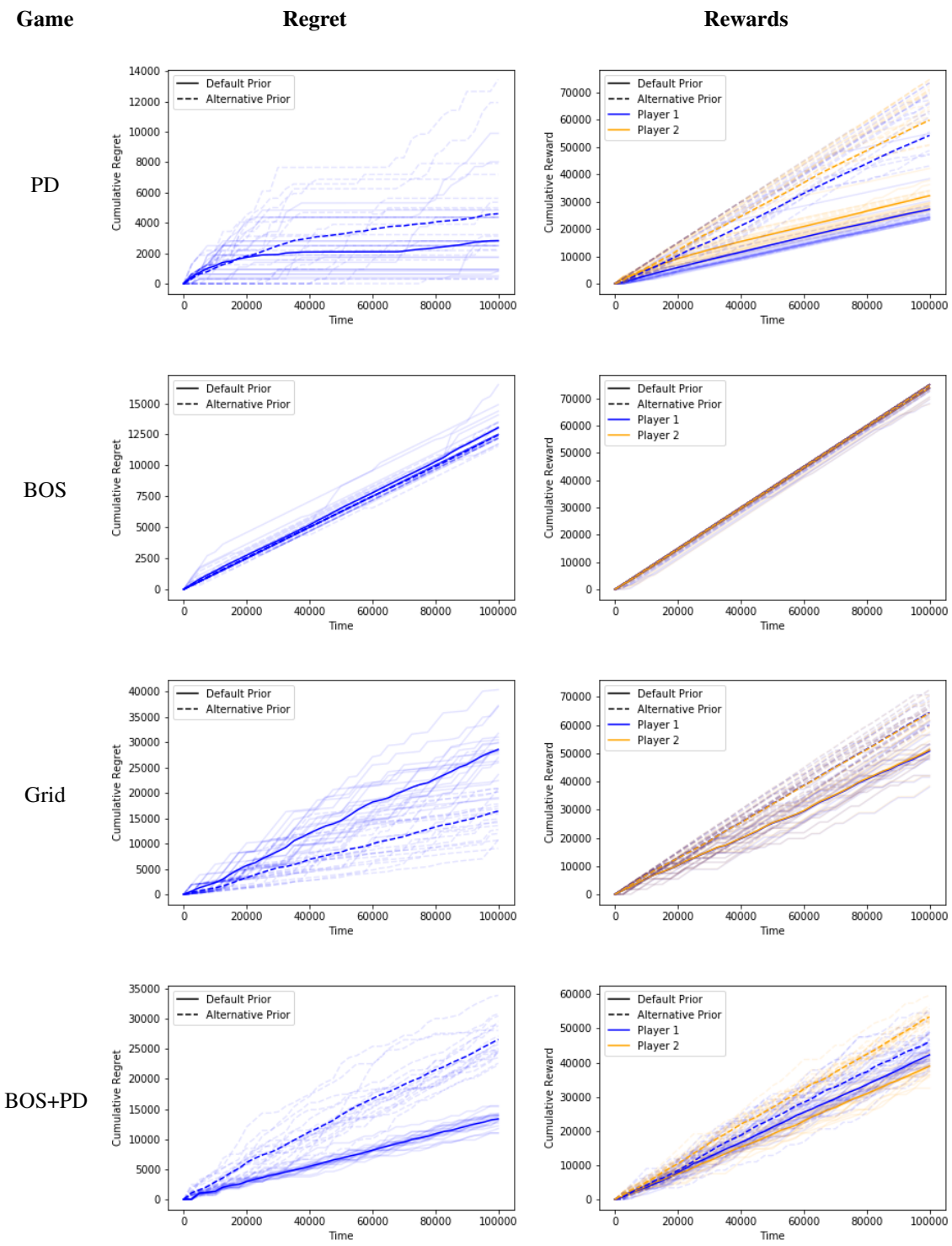


Figure 4: Results for self-play. Light curves are the 20 runs, and the bold curve is the pointwise average of the light curves.