Learning in Multi-Player Stochastic Games: Supplementary Material

William Brown[1]

[1]Computer Science Dept., Columbia University, New York, New York, USA

A OMITTED PROOFS

In Appendix A.1, we show hardness for the "adversarial MDP" problem. In Appendix A.2, we analyze the use of bandit algorithms for reaching correlated equilibria in games with stochastic rewards. In Appendix A.3, we introduce a general formulation of Bayesian games, for which obtain analogues of the results in Appendix A.2, which will be later used for analysis of learning in "single-controller" stochastic games. We prove our main results regarding PLL in Appendix A.4, and FastPLL in Appendix A.5. Our single-controller result is shown in Appendix A.6, and our "shared randomness" result for extending PLL is given in Appendix A.7.

A.1 PROOFS FOR SECTION 3: HARDNESS OF LEARNING IN ADVERSARIAL MDPS

Here we prove our hardness result for the finite-horizon or "episodic" adversarial MDP problem, where both transitions and rewards can change arbitrarily between episodes. We assume the adversary can pick the starting state as well, which is without loss of generality up to increasing the horizon by 1. This problem was shown to be at least as hard as learning parities with noise by Abbasi-Yadkori et al. [2013], and their reduction involves creating episodic MDPs with $H = \Theta(S)$. We strengthen this to NP-hardness, and for a horizon length of only 3. We do this by showing that the batch version of the problem, which we call the "multi-MDP", problem is at least as hard as 3-SAT, and as such is NP-hard to approximate within a factor of $\frac{7}{8} + \epsilon$, for any $\epsilon > 0$. By an online-to-batch reduction, this implies that there is no algorithm for the episodic adversarial MDP problem with poly-time per-round computation and $O(T^{1-\delta}\,\text{poly}(S))$ regret, for any $\delta > 0$, and for any dependence on N and H, unless NP $\subseteq$ BPP. Like Abbasi-Yadkori et al. [2013], our reduction only needs deterministic transitions, and so the hardness result also holds for the simpler "adversarial online shortest path problem".

A.1.1 The Offline Problem.

Consider the batch version of the adversarial MDP problem, which we call the "multi-MDP problem", where we are given a set of MDPs $\mathcal{M}$. Each MDP $M \in \mathcal{M}$ has a identical state and action spaces $\mathcal{X}$ and $\mathcal{A}$, as well as episode length H, but the transition and reward functions p and r can differ arbitrarily. Given $\mathcal{M}$ as input, the goal for the maximization problem is to output a single (possibly randomized and non-stationary) policy π which maximizes the average per-episode reward across all MDPs. The decision problem is to determine if any single policy achieves average reward at least R across $\mathcal{M}$. We assume that the per-episode reward in each MDP M is in $[0, d]$ for all policies, and that all instantaneous rewards are non-negative.

Theorem 1. *The decision version of the multi-MDP problem is* NP*-complete for horizon length $H \geq 3$. Further, the maximization version is* NP*-hard to approximate within a factor of $\frac{7}{8} + \epsilon$, for all $\epsilon > 0$.*

Proof. We reduce from 3-SAT. First we constrain ourselves to only considering deterministic policies. The idea is to encode each of the m clauses of a 3-SAT formula (on n variables) as set of six $(n+1)$-state MDPs. The states correspond to each of the variables as well as a "done" state, and the action space at each state is $\{0, 1\}$, corresponding to an assignment for the variable. Assume without loss of generality that the variables in the input formula are lexicographically ordered. Create one MDP for each of the six possible permutations of the literals in a clause; the episode will consist of three steps. For each of these MDPs, let the starting state s_i at step $h = 1$ correspond to the first literal x_i in the ordering. If x_i evaluates to True on input $\pi(s_i, 1)$ for a policy π, we transition to the "done" state, otherwise we transition to the state for the second literal s_j. Transitions proceed here accordingly for $\pi(s_j, 2)$ and likewise at the third state s_k for $\pi(s_k, 3)$. Once at the "done" state, we remain there until the end of the episode regardless of action. Transitioning to the "done" state from some other

state yields a reward of 1 and all other transitions yield a reward of 0.

If the input formula is satisfiable, then the stationary policy corresponding to the satisfying assignment will clearly obtain an average reward of 1. A non-stationary policy π defines six (not necessarily distinct) assignments of values to the n variables, for each permutation of the 3 timesteps. We can split the $6n$ MDPs into 6 sets, each of size n corresponding to one permutation, which are evaluated on the appropriate assignment of values. If the input formula is unsatisfiable, at least one MDP in each set will result in a reward of 0 Deciding whether any policy achieves an average reward of 1 or at most $1 - \frac{1}{n}$ is clearly in NP, as the best policy acts as a certificate, and so the problem is NP-complete.

This reduction also implies hardness of approximation. As is well-known, it is NP-hard to approximate MAX-3-SAT within a factor of $\frac{7}{8} + \epsilon$, for all $\epsilon > 0$. Suppose we can could compute a policy which obtains average reward at least $\frac{7}{8} + \epsilon$ in a set of MDPs with maximum possible average reward of 1. We can then apply to the above reduction to any input 3-SAT formula, resulting in a set of MDPs with a possible average reward of 1 if and only if the formula is satisfiable. If we can obtain average reward at least $\frac{7}{8} + \epsilon$ on this set, we must have average reward at least $\frac{7}{8} + \epsilon$ on the subset of MDPs corresponding to some permutation of literals. We can then extract an assignment from that permutation of timesteps in the policy which corresponds to an assignment which satisfies at least a $\frac{7}{8} + \epsilon$ fraction of the clauses in the input formula, implying the desired hardness result.

Any randomized policy can be derandomized without loss in average reward in polynomial time, implying that randomization does not help from a complexity perspective.

Lemma 1. *For any set of finite-horizon MDPs, any randomized policy can be converted to a deterministic non-stationary policy in polynomial time without decreasing average reward.*

Proof. Consider the uniform distribution over MDPs in the set $\mathcal{M}$ and the induced distribution over states in the final timestep. By the Markov property and the assumption of a fixed policy, the conditional distribution of actions at a state is independent of the MDP as well as the sequence of states visited. Each action with positive support has some expected reward when taking the expectation over the MDP distribution, transitions, and previous action selections; playing the maximum action at each state does not decrease expected reward. We can apply this to each previous step by backward induction, as downstream conditional expected values for actions at each state are still defined, giving us a fully deterministic policy. □

As such, the hardness result holds even for algorithms which output randomized non-stationary policies.

□

A.1.2 Hardness for Regret Minimization and Black-Box NFCCEs in Stochastic Games

We use Theorem 9 to prove our hardness result for quickly vanishing regret in the adversarial MDP problem.

Restatement of Theorem 2. *Assuming* NP $\not\subseteq$ BPP*, there is no algorithm with polynomial time per-round computation which has $O(T^{1-\delta}\text{poly} \cdot (S))$ regret algorithm for the adversarial MDP problem with $H \geq 3$, for any $\delta > 0$.*

Proof. By the standard online-to-batch reduction from Cesa-Bianchi et al. [2004], we can convert an algorithm with small regret to an algorithm for MAX-3-SAT. Suppose we had an algorithm with regret $O(T^{1-\delta}S^k)$ for constants k and δ. Take $T \gg S^{k/\delta}$ but still polynomial in S such that the average regret is $o(1)$. Apply the reduction from Theorem 9 to a 3-SAT instance on S variables and then run the algorithm for T steps, sampling from the uniform distribution over the constructed MDPs at each episode. By the main result (Theorem 4) from Cesa-Bianchi et al. [2004], the empirically optimal policy over the historical sequence achieves a value within $o(1)$ of the optimum with high probability. This would imply a polynomial time algorithm which beats a $\frac{7}{8} + \epsilon$ approximation for MAX-3-SAT, which is impossible unless NP $\subseteq$ BPP due to Håstad [1997]. □

This directly implies Corollary 2.1, where the horizon is increased to 4 to account for the starting state in a finite-horizon stochastic game being random rather than adversarial (in our reduction, one can add a "starting state" from which the adversary selects the next state).

A.2 GAMES WITH STOCHASTIC REWARDS

Recall that for a game with stochastic rewards, we consider all players running an adversarial bandit algorithm $\mathcal{B}$ (such as SR-MAB). A step in our analysis introduces an additional $\log(1/\epsilon)$ term beyond the runtime of SR-MAB for target average regret ϵ, yet with less dependence on N. This is not an issue if N is sufficiently large as a function of ϵ, but if this is not the case we extend the runtime to that which would be required if $N = \Omega\left(\sqrt[3]{\frac{\log(1/\epsilon)}{\log(\log(1/\epsilon))}}\right)$, which can only increase average regret; we denote this runtime function by $B(\epsilon, N)$.

Theorem 3. *When players in a game with stochastic rewards x select actions using $\mathcal{B}$ for $T \geq B(\epsilon/4, N)$ rounds, the sequence of action profiles is an ϵ-correlated equilibrium for the game, where the expectation is taken with respect to the tensor distribution as well as $\mathcal{B}$.*

Proof of Theorem 3. We begin with a lemma relating the runtime of SR-MAB to the term which we will use in our martingale analysis of the "sampling error" of the realized sequence of reward tensors versus the average tensor $\bar{\theta}$.

Lemma 2. *If* $N = \Omega\left(\sqrt[3]{\frac{\log(1/\epsilon)}{\log(\log(1/\epsilon))}}\right)$ *then* $\frac{N^3 \log(N)}{\epsilon^2} = \Omega\left(\frac{N\log(N)+\log(1/\epsilon)}{\epsilon^2}\right)$.

Proof of Lemma 2. It suffices to show that $N^3\log(N) = \Omega(\log(1/\epsilon))$. Plugging in our expression for N, we have that

$$\begin{aligned} N^3 \log N &= \Theta(N^3 \log(N^3)) \\ &= \Omega\left(\frac{\log(\frac{1}{\epsilon}) \cdot \left(\log\log(\frac{1}{\epsilon}) - \log\log\log(\frac{1}{\epsilon})\right)}{\log\log(\frac{1}{\epsilon})}\right) \\ &= \Omega(\log(1/\epsilon)). \end{aligned}$$

□

By the regret guarantee of $\mathcal{B}$, each player has expected average swap regret at most $\epsilon/4$ with respect to the sampled sequence of reward tensors $\{\theta^t\}_{t\in[T]}$, which we denote $\overline{\text{Reg}}_{\mathcal{B}}^{\{\theta^t\}}$. For a player i, consider some swap function f. Let $X_f^t = u_i(f(a_i^t); a_{-i}^t, \theta^t) - u_i(f(a_i^t); a_i^t, \bar{\theta})$ for an action profile and tensor (a^t, θ^t), i.e. the difference between this player's reward from using f on θ^t versus the average tensor $\bar{\theta}$, given the action profile a^t. Let $Y_f^t = \sum_{j=1}^t X_f^t$. For a distribution over tensors, and any sequence of action profiles where a_t is independent of θ^t given actions and tensors for $1,\ldots,t-1$, the sequence $Y_f^1,\ldots,Y_f^t$ is a martingale with respect to the sequence X_i^f. To see this, note that for any fixed a^t, X_f^t is in $[-1,1]$ as rewards are in $[0,1]$, and $\mathbb{E}[Y_f^t \mid X_f^1,\ldots,X_f^{t-1}] = Y_f^{t-1}$, as $\mathbb{E}[X_f^t \mid X_f^1,\ldots,X_f^{t-1}] = 0$ by the definition of $\bar{\theta}$.

Let $T \geq \frac{32(N\log(N)+\log(8/\epsilon))}{\epsilon^2} = \frac{32\log(8N^N/\epsilon)}{\epsilon^2}$ by Lemma 2. By the Azuma-Hoeffding inequality we have that

$$\begin{aligned} \Pr[\left|Y_f^T\right| \geq \frac{\epsilon T}{4}] &\leq 2\exp\left(\frac{-\epsilon^2 T}{32}\right) \\ &\leq \frac{\epsilon}{4N^N}. \end{aligned}$$

Union-bounding over all $f \in \mathcal{F}$, we then have that

$$\max_{f\in\mathcal{F}} \left|\frac{1}{T}\sum_{t=1}^T u_i(f(a_i^t); a_{-i}^t, \theta^t) - u_i(f(a_i^t); a_i^t, \bar{\theta})\right| \leq \frac{\epsilon}{4}$$

with probability at least $1-\frac{\epsilon}{4}$. As such, the average utility of a swap function on the sequence deviates from its expected utility on the distribution by at most $\epsilon/4$ with probability at least $\epsilon/4$, holding simultaneously for all functions, including the identity function I (our benchmark for swap regret). As such, with probability $1-\epsilon/4$, the difference in swap regret on the sequence and the distribution, denoted by $\left|\overline{\text{Reg}}_{\mathcal{B}}^{\bar{\theta}} - \overline{\text{Reg}}_{\mathcal{B}}^{\{\theta_t\}}\right|$, is at most $\epsilon/2$. Using the maximal deviation of 1 as a bound for the difference for the remaining probability, we then have that

$$\mathbb{E}\left[\left|\overline{\text{Reg}}_{\mathcal{B}}^{\bar{\theta}} - \text{Reg}_{\mathcal{B}}^{\{\theta^t\}}\right|\right] \leq (1-\epsilon/4)\cdot\epsilon/2 + \epsilon/4 \leq \frac{3\epsilon}{4}.$$

Therefore by our bound on $\mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B}}^{\{\theta^t\}}\right]$ and linearity of expectation:

$$\begin{aligned} \mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B}}^{\bar{\theta}}\right] &= \mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B}}^{\{\theta^t\}}\right] + \mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B}}^{\bar{\theta}} - \overline{\text{Reg}}_{\mathcal{B}}^{\{\theta^t\}}\right] \\ &\leq \mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B}}^{\{\theta^t\}}\right] + \mathbb{E}\left[\left|\overline{\text{Reg}}_{\mathcal{B}}^{\bar{\theta}} - \overline{\text{Reg}}_{\mathcal{B}}^{\{\theta^t\}}\right|\right] \\ &\leq \epsilon. \end{aligned}$$

As no player can improve average utility in expectation for $\bar{\theta}$ by more than ϵ with any swap function, the uniform distribution over the sequence of action profiles is an ϵ-correlated equilibrium for x when taking the expectation over both the profile sequence and the generating process using $\mathcal{B}$ and samples of reward tensors. □

Corollary 1.1 (Restatement of Corollary 3.1)**.** *When all agents in a game with stochastic rewards x play according to $\mathcal{B}$ for at least $\frac{2\log(5M/\delta)}{\eta^2}\cdot B(\epsilon/8, N)$ rounds, simultaneously restarting $\mathcal{B}$ every $B(\epsilon/8, N)$ rounds, the resulting sequence of actions is an $(\epsilon/2+\eta/2)$-correlated equilibrium for x with probability at least $1-\delta/5$.*

Further, let $V_i^{\mathcal{B}}(x) = \mathbb{E}_{\mathcal{B},r}\left[\frac{1}{T}\sum_{t=1}^T u_i(a_i^t; a_{-i}^t, \theta^t)\right]$ and let $\hat{V}_i^{\mathcal{B}}(x)$ be the average utility received by player i over all rounds. With probability at least $1-2\delta/5$, $\left|V_i^{\mathcal{B}}(x) - \hat{V}_i^{\mathcal{B}}(x)\right| \leq \eta/2$ simultaneously for all players.

Additionally, the computed estimate is within η of player i's expected average reward for playing the game according to the resulting policy distribution with probability at least $1-2\delta/5$.

Proof of Corollary 3.1. The swap regret of a sequence is upper-bounded by the sum of the swap regret values of a uniform partition of the sequence, as the latter may use a different swap function on each sequence while the former is restricted to only using a single function. As such, we can bound the average regret of our sequence by averaging the average swap regret values between restarts.

Both average utility and average swap regret (with respect to x) over $B(\epsilon/8)$ are random variables taking values in $[0,1]$, and the mean of the latter is at most $\epsilon/2$ by Theorem 3. Recall from the proof of Theorem 3 that the expected average reward deviation of the identity function on the sequence and distribution has mean zero (by nature of it being a martingale), and it takes values in $[-1,1]$. The result

then follows from applying Hoeffding's inequality to the average of the samples we receive of the random variables, bounding deviation by $\eta/2$ (or η), and union-bounding over all players and failure probabilities.

□

A.3 CORRELATED EQUILIBRIA IN BAYESIAN GAMES

We also give a convergence result for learning in Bayesian games. The Bayesian game formulation we consider is quite general (in particular, we remove the "independent private value" assumption from the model considered in Hartline et al. [2015], and allow signals and rewards to be arbitrarily correlated across players), and can be viewed as a partial-information generalization of games with stochastic rewards. When all players use our described method, the sequence of policy profiles played by all players converges to an approximate Bayes correlated equilibrium in polynomial time.

Definition 1 (Bayesian Games)**.** *A Bayesian game $y = (\mathcal{A}, M, \psi, p, r, u)$ has M players and is specified by a set of action profiles $\mathcal{A} = \times_{i\in[M]} \mathcal{A}_i$, a signal function $\psi : \mathcal{X} \to \Psi$ where $\Psi = \times_{i\in[M]}\Psi_i$, and a distribution over states $p \in \Delta(\mathcal{X})$. Each state x denotes a game with stochastic rewards, with its distribution over reward tensors given by $r : \mathcal{X} \to \Delta(\Theta)$. Players' utilities, given by $u : \mathcal{A} \times \Theta \to [0, 1]^N$, depend on the realization of $\theta \sim r(x)$. Players only observe a signal of the state $\psi_i(x)$, and never observe θ or x directly.*

We assume that $|\mathcal{A}_i| = N$ for all agents, and we will let $S_i = |\Psi_i|$ and $S = \max_i S_i$. In this model of a Bayesian game, a state x is drawn from p, each agent i observes a signal $\psi_i(x)$ and selects an action a_i, then receives utility $u_i(a_i; a_{-i}, \theta)$, where θ is drawn from $r(x)$. We note that Bayesian games are often defined in such a way where states and reward tensors are treated as equivalent. This formulation of a Bayesian game is similar to the "information set" model often considered in partially-observable Markov decision processes and extensive-form games. However, our result for Bayesian games will not depend on the size of $\mathcal{X}$ or Θ. Here, one could treat $\mathcal{X}$ and Θ as identical, but we maintain the distinction for continuity in exposition with our sections on stochastic games. It is without loss of generality that we assume u depends only on a and θ, not x, as we can encode arbitary distributions over reward vectors in $[0, 1]^M$ for each state with a distribution over reward tensors.

The definition of correlated equilibrium in Bayesian games given in Bergemann and Morris [2016] refers to a *decision rule*, given by a distribution over action profile recommendations for each state and set of types, which is *obedient* in the sense that no player can improve by deviating from the recommendations for any action-type pair. The method we present here will converge to a joint distribution over policy profiles, denoting an action recommendation for each signal, which will be independent of the state and reward tensor distributions, and which satisfies this definition of Bayes correlated equilibrium. Several other definitions are considered in the literature as well Forges [1993].

We are aware of only one paper, Hartline et al. [2015], which considers learning correlated equilibria in Bayesian games through the lens of polynomial time convergence, where the primary focus is on analyzing the Price of Anarchy and connections to learning in auctions. They consider the *independent private value model* of Bayesian games, There, the assumption is made that players have "types" which fully characterize their rewards for any action profile, and further that these types are drawn from a product distribution. In their approach, each agent runs parallel copies of a no-regret algorithm for each type, and actions are sampled from each algorithm every round, which they interpret as the sampling of a strategy mapping types to actions. Our model is a generalization of this setting, as we allow types (signals) to be arbitrarily correlated with eachother as well as with the reward tensors. To our knowledge, the approach we give here is the first which converges to a Bayes correlated equilibrium in polynomial time for such a general formulation of Bayesian games.

Here will consider *policies* $\pi_i : \Psi_i \to \mathcal{A}_i$ for an agent i, with $\pi_i \in \Pi_i$ and $\Pi = \times_{i\in[M]}\Pi$, which are functions mapping their signals to actions. In our setting, a Bayes correlated equilibrium is a distribution over policy profiles such that no agent can benefit by deviating from policy recommendations.

Definition 2 (Bayes Correlated Equilibria)**.** *A Bayes correlated equilibrium for a Bayesian game is a distribution over policy profiles given by $D \in \Delta(\Pi)$ such that for all players i and all swap functions $f \in \mathcal{F}^{\Psi_i} : \mathcal{A}_i \times \Psi_i \to \mathcal{A}_i$,*

$$\mathop{\mathbb{E}}_{\pi\sim D,\theta\sim r(x),x\sim p}[U_i] \geq \mathop{\mathbb{E}}_{a\sim D(x),\theta\sim r(x),x\sim p}[U_i^f],$$

with $U_i = u_i(\pi_i(\psi_i(x)); a_{-i}, \theta)$ and $U_i^f = u_i(f(\pi_i(\psi_i(x)), \psi_i(x)); a_{-i}, \theta)$, where a_{-i} is the vector of actions $[\pi_j(\psi_j(x))]$ for agents $j \neq i$, and where the policy vector π is sampled independently from x. Such a distribution is an ϵ-Bayes correlated equilibrium if for all players and swap functions,

$$\mathop{\mathbb{E}}_{\pi\sim D,\theta\sim r(x),x\sim p}[U_i] \geq \mathop{\mathbb{E}}_{a\sim D(x),\theta\sim r(x),x\sim p}[U_i^f] - \epsilon.$$

The smallest quantity ϵ for which the above holds for agent i is their average $\mathcal{F}^{\Psi_i}$-regret for a policy distribution.

We let $\mathcal{B}_S$ denote the *parallel bandit* algorithm consisting of S copies of $\mathcal{B}$, with one copy for each type. At the beginning of each round, agents sample actions from each copy of $\mathcal{B}$,

thereby creating a policy π_i^t for the round. Upon observing their signal ψ_i^t, they play the action $\pi_i^t(\psi)i^t)$, update the copy of $\mathcal{B}$ corresponding to ψ_i with their observed reward, and record a record a reward of 0 for all other copies. We show that when agents play according to $\mathcal{B}_S$, the sequence of policies converges to an approximate equilibrium for the Bayesian game.

Theorem 2. *When players in a Bayesian game y select actions using $\mathcal{B}_S$ for $T \geq B(\frac{\epsilon}{4S}, N)$ rounds, where the state is sampled independently each round and the reward tensor is sampled from that state's distribution, the sequence of policies is an ϵ-correlated equilibrium for the game, where the expectation is taken with respect to the state, tensor, and action profile distribution as well as the randomness of $\mathcal{B}$.*

Proof of Theorem 2. The proof is quite similar to that for Theorem 3. We bound the expected average swap regret for each copy of B by ϵ/S, which then bounds the total average swap regret (with respect to the policy class) by ϵ.

By the guarantee of the algorithm $\mathcal{B}$, each player's copy of $\mathcal{B}$ for a signal ψ_i has expected swap regret at most $\frac{\epsilon}{4S}$ with respect to the sampled sequence of states and reward tensors (where rewards are 0 when the corresponding signal is not observed), which we denote $\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}$. The average swap regret for the entire sequence will be the sum of the swap regrets for each signal, denoted $\overline{\text{Reg}}_{\mathcal{B}_S}^{\{\theta^t,x^t\}} = \sum_{\psi_i \in \Psi_i} \overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}$, as the deviations considered by the function class $\mathcal{F}^{\Psi_i}$ are equivalent to choosing any $f \in \mathcal{F}$ for each signal.

For a player i and signal ψ_i, upon fixing the vector of opponent policies π_{-i}, there is some fixed expected reward for each action, conditional on observing ψ_i, given by:

$$\bar{\theta}(a_i; \psi_i, \pi_{-i}) = \mathop{\mathbb{E}}_{x\sim p, \theta \sim r(x)} \left[U_i^{\psi} \times \mathbf{1}[\psi_i(x) = \psi_i] \right],$$

where $U_i^{\psi} = u_i(a_i; (\pi_j(\psi_j(x)))_{j\neq i}, \theta)$. In round t of the game, the reward that player i's copy of $\mathcal{B}$ associated with ψ_i will receive for playing action a_i is a random variable in $[0,1]$ with mean $\bar{\theta}(a_i; \psi_i, \pi_{-i})$, where we view π_{-i} as being fixed prior to the realization of x and θ. The regret bound for that copy of $\mathcal{B}$ holds for the realized sequence of vectors (determined by π_{-i}^t, x^t, and θ^t) of these rewards for all actions a_i. We will be interested in bounding the average reward deviation of swap functions between this sequence and the sequence $((\bar{\theta}(a_i; \psi_i, \pi_{-i}^t))_{a_i \in \mathcal{A}_i})_{t\in[T]}$.

Consider some swap function $f \in \mathcal{F}$. We can again define a martingale which tracks the deviation of the performance of f on the sampled sequence versus the underlying game distribution. Let $X_{f,\psi_i}^t = (u_i(f(a_i^t); (\pi_j^t(\psi_j(x^t)))_{j\neq i}, \theta^t) \cdot \mathbf{1}[\psi_i(x^t) = \psi_i] - \bar{\theta}(f(a_i^t); \psi_i, \pi_{-i}^t))$ for a policy profile, signal, and tensor $(\pi_{-i}^t, \psi_i^t, \theta^t)$, i.e. the difference between this player's observed and expected reward from using f with the copy of $\mathcal{B}$ associated with ψ_i, given opponent policies π_{-i}^t and their own sampled action a_i^t for signal ψ_i. Let $Y_{f,\psi_i}^t = \sum_{j=1}^t X_{f,\psi_i}^t$. For a distribution over states and tensors, and any sequence of action profiles where a_t is independent of θ^t given actions and tensors for $1, ..., t-1$, the sequence $Y_{f,\psi_i}^1, ..., Y_{\psi_i}^t$ is a martingale with respect to the sequence X_{f,ψ_i}^t. To see this, note that for any fixed a^t, X_{f,ψ_i}^t is in $[-1,1]$ as rewards are in $[0,1]$, and $\mathbb{E}[Y_{f,\psi_i}^t \mid X_{f,\psi_i}^1, \ldots, X_{f,\psi_i}^{t-1}] = Y_{f,\psi_i}^{t-1}$, as $\mathbb{E}[X_{f,\psi_i}^t \mid X_{f,\psi_i}^1, \ldots, X_{f,\psi_i}^{t-1}] = 0$ by the definition of $\bar{\theta}$.

Let $T \geq \frac{32S^2(N\log(N)+\log(8S/\epsilon))}{\epsilon^2} = \frac{32S^2\log(8SN^N/\epsilon)}{\epsilon^2}$ by Lemma 2. By the Azuma-Hoeffding inequality we have that

$$\begin{aligned}\Pr[\left|Y_{f,\psi_i}^T\right| \geq \frac{\epsilon T}{4S}] &\leq 2\exp\left(\frac{-\epsilon^2 T}{32S^2}\right) \\ &\leq \frac{\epsilon}{4SN^N}.\end{aligned}$$

Union-bounding over all $f \in \mathcal{F}$, we then have that

$$\max_{f\in\mathcal{F}} \left| \sum_{t=1}^T \frac{U_i^{\psi,t} \cdot \mathbf{1}[\psi_i(x^t) = \psi_i] - \bar{\theta}(f(a_i^t); \psi_i, \pi_{-i}^t)}{T} \right| \leq \frac{\epsilon}{4S}$$

where $U_i^{\psi,t} = u_i(f(a_i^t); (\pi_j^t(\psi_j(x^t)))_{j\neq i}, \theta^t)$ with probability at least $1 - \frac{\epsilon}{4S}$. As such, the average utility of a swap function on the sequence applied to the copy of $\mathcal{B}$ for ψ_i deviates from its expected utility on the distribution by at most $\frac{\epsilon}{4S}$ with probability at least $\frac{\epsilon}{4S}$, holding simultaneously for all functions in $\mathcal{F}$, including the identity function I (our benchmark for swap regret). As such, with probability $1 - \frac{\epsilon}{4S}$, the difference in average swap regret on the sequence and the distribution for this $\mathcal{B}$ copy, denoted by $\left|\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\bar{\theta}} - \overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}\right|$, is at most $\frac{\epsilon}{2S}$. Using the maximal deviation of 1 as a bound for the difference for the remaining probability, we then have that

$$\mathbb{E}\left[\left|\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\bar{\theta}} - \text{Reg}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}\right|\right] \leq (1 - \frac{\epsilon}{4S}) \cdot \frac{\epsilon}{2S} + \frac{\epsilon}{4S} \leq \frac{3\epsilon}{4}.$$

Therefore by our bound on $\mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B}}^{\{\theta^t,x^t\}}\right]$ and linearity of expectation:

$$\begin{aligned}\mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\bar{\theta}}\right] &= \mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}\right] + \mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\bar{\theta}} - \overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}\right] \\ &\leq \mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}\right] + \mathbb{E}\left[\left|\overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\bar{\theta}} - \overline{\text{Reg}}_{\mathcal{B},\psi_i}^{\{\theta^t,x^t\}}\right|\right] \\ &\leq \epsilon.\end{aligned}$$

Summing over each copy of $\mathcal{B}$ gives us that $\mathbb{E}\left[\overline{\text{Reg}}_{\mathcal{B}_S}^{\{\theta^t,x^t\}}\right] \leq \epsilon$, as average swap regret (with respect to $\mathcal{F}^{\Psi_i}$) for the distribution can be decomposed into swap regret for each signal (with respect to $\mathcal{F}$) just as for the sequence of states and tensors. As no player can improve average utility in expectation for $\bar{\theta}$ by more than

ϵ with any swap function $\mathcal{F}^{\Psi_i}$, the uniform distribution over the sequence of policy profiles is an ϵ-correlated equilibrium for y when taking the expectation over both the profile sequence and the generating process using $\mathcal{B}_S$ and samples of states and reward tensors. □

Again, if desired we can simultaneously obtain an accurate estimate of the value $V_i^{\mathcal{B}_S}(y)$ of this equilibrium-generating process for each player, and boost regret bounds to high probability, with repeated restarts.

A.4 ANALYSIS FOR PLL

Showing Theorem 4 for BILL is straightforward and a proof can be obtained by simplifying the analysis of PLL in Theorem 5. We restate the description of PLL here, with explicit constants for the terms whose asymptotic descriptions were given in the body.

Algorithm 2: Parallel Local Learning. Initialize $\hat{V}_i^{\mathcal{B}}(x,h) = H - h + 1$ for each pair (x,h), as well as a visit counter $c(x,h)$ for each pair set to 0. Let $W = \max(W_1, W_2)$, where $W_1 = \frac{128S^4H^6\log(2S/\delta')}{\epsilon^2}$, $W_2 = \frac{512H^4\log(5M/\delta')}{\epsilon^2}$, and $\delta' = \frac{\epsilon\delta}{192SH^4((S+1)^H+1)\cdot\max(S,4H^7/\epsilon)}$. Let $L \geq \max\left(\frac{64S^2H^3WB}{\epsilon}, \frac{256SH^4WB}{\epsilon^2}\right)$. Initialize a copy of $\mathcal{B}$ at each pair, specified to run for $B = B(\frac{\epsilon}{16H}, N)$ steps. Until termination, run the following procedure for each epoch:

- Run for L trajectories, using $\mathcal{B}$ at each pair, counting rounds and updating actions for a copy of $\mathcal{B}$ only when the corresponding pair is visited. Record rewards as the sum of the observed reward as well as the value estimate for the next pair visited in that trajectory, scaled to [0,1].
- Consider the last step $h \in [H]$ where an unlocked pair's counter crossed $\frac{16H^2WB}{\epsilon}$ in the epoch. Lock all unlocked states at this step with appropriate estimates which were previously unlocked, compute value estimates $\hat{V}_i^{\mathcal{B}}(x,h)$ as the average reward over the corresponding $\frac{16H^2WB}{\epsilon}$ visits, then reset all copies of $\mathcal{B}$, counters, and estimates at *earlier* pairs $(h' < h)$.
- Terminate if no pair's counter crosses $\frac{16H^2WB}{\epsilon}$ in the epoch.

Restatement of Theorem 5. *PLL terminates after at most $(S+1)^H + 1$ epochs. After termination, for each pair (x, h), consider the uniform distribution over action profiles $D(x,h)$ played since that pair was last reset. Let D be the distribution over policy profiles where the action profile for each pair (x,h) is sampled independently from $D(x,h)$. With probability at least $1-\delta$, D is an ϵ-EFCE for the game.*

Proof of Theorem 5. We first give a worst-case bound on the runtime, then proceed with our analysis of the regret of the resulting action profile distributions. At termination, for any pair (x,h) with no visits since it was last reset, we can let the distribution $D(x,h)$ over action profiles be arbitrary for the purposes of our our analysis.

Lemma 3 *PLL runs for at least H epochs, and at most $(S+1)^H + 1$ epochs.*

Proof. All pairs start unlocked, and some pair in each step is visited at least $\frac{16H^2WB}{\epsilon}$ per epoch by pigeonhole, so the algorithm will not terminate unless there is a locked pair for every step. States are only moved from unlocked to locked at one step per epoch, and so there must be at most H epochs to lock some pair in all steps.

We can bound the number of epochs by bounding the number of epochs in which a pair at some step can become locked. Observe that a locked pair at step H will only become locked in one epoch and will never become unlocked afterwards. A pair at step $H-1$ will become locked in at most S epochs, as it will only become locked after at least one pair at step H is locked, and then can be unlocked at most $S-1$ times for the remaining unlocked pairs at step H. In general, the number of epochs in which a state can become locked is bounded by the number of epochs in which a downstream state can become locked. Let $g(h)$ denote this bound on the number of epochs in which a pair at step h can be locked, which is given by:

$$\begin{aligned} g(h) &= \sum_{i=h+1}^{H} Sg(i) \\ &= Sg(h+1) + \sum_{i=h+2}^{H} Sg(i) \\ &= (S+1)g(h+1) \\ &= (S+1)^{H-h}g(H) \\ &= (S+1)^{H-h}, \end{aligned}$$

as $g(H) = 1$. The total number of epochs before termination is then bounded by

$$1 + \sum_{i=1}^{H} Sg(i) = g(0) + 1 = (S+1)^H + 1,$$

accounting for the last epoch in which no states are locked. □

We now show that each agent has small regret with respect to $\mathcal{F}$ under the resulting policy distribution D with high probability, which coincides with the definition of extensive-form correlated equilibria we consider, as D is a product distribution across pairs. An important object in this analysis is the expected distribution over state visitations when

players use $\mathcal{B}$ at each pair with a fixed set of values. Just as there is some fixed distribution over average rewards when players play $\mathcal{B}$ in a game for many rounds, there is also a fixed distribution over transitions when using $\mathcal{B}$ at a pair in a stochastic game, given fixed sets of value estimates for downstream states.

When all agents use a bandit algorithm $\mathcal{B}$ at a pair $(x, h-1)$ for B trajectories *where* $(x, h-1)$ *is visited*, augmenting rewards with downstream value estimates $\hat{V}_i(x', h)$ for each player i and state x', there is some expected proportion of those trajectories that each state will be visited at step h, which we denote by:

$$p_{\hat{V}}(x'; x, h) = \mathop{\mathbb{E}}_{\mathcal{B}, p(h)} \left[\frac{1}{B} \sum_{t=1}^{B} \mathbf{1} \left[\tau(a^t, x) = x' \right] \right]$$

We can also define the probability that a pair is visited in a trajectory, assuming that the distribution of transitions between pairs is given by $p_{\hat{V}}$, which we denote by $q_{\hat{V}}(\cdot, \cdot)$:

$$q_{\hat{V}}(x, 1) = p_0(x),$$
$$q_{\hat{V}}(x, h) = \sum_{x' \in \mathcal{X}} q_{\hat{V}}(x', h-1) \cdot p_{\hat{V}}(x; x', h-1).$$

For a distribution $D(x, h)$ of action profiles for each pair, we can also define transition probabilities between pairs in a trajectory when action profiles are selected independently for each pair:

$$p_D(x'; x, h) = \Pr_{D(x,h-1), p(h)} \left[\tau(a, x) = x' \right],$$

as well as expected visitation frequencies for each pair in a trajectory:

$$q_D(x, 1) = p_0(x),$$
$$q_D(x, h) = \sum_{x' \in \mathcal{X}} q_D(x', h-1) \cdot p_D(x; x', h-1).$$

If a pair (x, h) is visited sufficiently often with fixed downstream values $\hat{V}$, then both the empirical transition distribution and the transition distribution *when transition functions are resampled* are close to $p_{\hat{V}}(x, h)$.

We prove a lemma about the composition of bounds on the total variation distance in this setting.

Lemma 3. *For distribution functions p and $\hat{p}$ mapping $\mathcal{X} \times [H]$ to $\Delta(\mathcal{X})$, and q and $\hat{q}$ mapping $[H]$ to $\Delta(X)$, where $q(x, h+1) = \sum_{x' \in \mathcal{X}} q(x', h) \cdot p(x; x', h)$ and $q(x, 1)$ can be arbitrary (and with $\hat{q}$ defined likewise with respect to $\hat{p}$), then with $d_q^{h+1} = d_{TV}(q(\cdot, h+1), \hat{q}(\cdot, h+1))$,*

$$d_q^{h+1} \leq d_{TV}(q(\cdot, h+1), \hat{q}(\cdot, h+1)) + \sum_{x' \in \mathcal{X}} q(x', h) \cdot d_{TV}(p(\cdot; x', h), \hat{p}(\cdot; x', h)).$$

Proof of Lemma 3.

$$\begin{aligned} d_q^{h+1} &= \frac{1}{2} \sum_{x \in \mathcal{X}} \left| \sum_{x' \in X} q(x', h) p(x; x', h) - \hat{q}(x', h) \hat{p}(x; x', h) \right| \\ &\leq \frac{1}{2} \sum_{x' \in X} \left(q(x', h) \sum_{x \in \mathcal{X}} |p(x; x', h) - \hat{p}(x; x', h)| \right) \\ &\quad + \sum_{x' \in X} \left(|q(x', h) - \hat{q}(x', h)| \sum_{x \in \mathcal{X}} p_D(x; x', h) \right) \\ &= d_{TV}(q(\cdot, h), \hat{q}(\cdot, h)) \\ &\quad + \sum_{x' \in X} q(x', h) \cdot d_{TV}(p(\cdot; x', h), \hat{p}(\cdot; x', h)). \end{aligned}$$

□

In Lemma 4 we show that in each epoch, for any pair where $q_{\hat{V}}(x, h)$ is sufficiently large (for the estimates $\hat{V}$ used in that epoch), the number of times in that epoch (x, h) is visited is close to expectation. We then show that any state which is unlocked at termination will almost surely be visited infrequently when agents play according to $D(x, h)$ at each state.

Lemma 4. *In any epoch where current value estimates are given by $\hat{V}$ for each player and pair, with probability at least $1 - \frac{\delta}{3((S+1)^H+1)}$, every pair (x, h) where $q_{\hat{V}}(x, h) \geq \frac{\epsilon}{8SH^2}$ reaches the locking threshold by the completion of the epoch.*

Proof. We proceed by showing that in each epoch, with high probability, the total variation distance between $q_{\hat{V}}(\cdot, h)$ and the empirical distribution over visited states at step h is small for every h. We prove this inductively.

Consider a sequence of BW visits to a pair (x, h), where W runs of $\mathcal{B}$ are completed. For each run of $\mathcal{B}$, the number of visits to a given pair $(x', h+1)$ is a random variable in $[0, B]$ with mean $B \cdot p_{\hat{V}}(x'; x, h)$, determined by the randomness of each player's copy of $\mathcal{B}$ as well as the game. For such a pair $(x', h+1)$, let X_i denote the $[0, 1]$ scaling of this random variable for the ith of the W runs, which has mean $p_{\hat{V}}(x'; x, h)$, and let $X = \sum_{i=1}^{W} X_i$. Each run is independent and so by Hoeffding's inequality,

$$\Pr \left[|X - \mathbb{E}[X]| \leq \frac{2W\epsilon'}{S} \right] \leq 2 \exp\left(-8W(\epsilon')^2/S^2\right)$$

which is at most $\frac{\delta'}{S}$ if $W \geq \frac{S^2 \log(2S/\delta')}{8(\epsilon')^2}$. This holds for all states x' with probability $1 - \delta'$ by a union bound, at which point we have that the empirical visitation frequency for every state x' is within $\pm 2\epsilon'/S$ of $p_{\hat{V}}(x'; x, h)$, implying that the total variation distance is at most ϵ'.

Let $\epsilon' = \frac{\epsilon}{32SH^3}$. We have that $W \geq W_1 = \frac{128 S^4 H^6 \log(2S/\delta')}{\epsilon^2}$, and the empirical transition distribution for a window of BW steps at a state (x, h) has

total variation distance with $q_{\hat{V}}(\cdot; x, h)$ at most $\frac{\epsilon}{32SH^3}$ with probability at least $1 - \delta'$. Recall that $\delta' \leq \frac{\epsilon\delta}{192SH^4((S+1)^H+1)\cdot\max(S,4H^7/\epsilon)} = \frac{\delta BW}{3LH((S+1)^H+1)}$; there are LH total *steps* in each epoch, which fall into at most $\frac{LH}{BW}$ completed windows of length BW, and so the above holds for all windows in an epoch with probability at least $1 - \frac{\delta}{3((S+1)^H+1)}$ by a union bound. The bound then holds for every pair and epoch with probability at least $1 - \delta/3$.

Using bounds on the empirical outgoing visitation distributions for each pair which is visited sufficiently often, we can obtain a bound on the total variation distance between $q_{\hat{V}}$ and the empirical visitation distribution over the epoch at each step, by Lemma 3. All but at most $2SWB$ of the steps fall into separate but contiguous windows of length WB, as there can be at most two "incomplete" windows (at the start and end) for each state where we cannot apply the above analysis. Observe that accounting for these unfinished windows increases the total variation distance between $q_{\hat{V}}$ and the empirical visitation distribution by at most $\frac{\epsilon}{32SH^3}$ if $\frac{2SWB}{L} \leq \frac{\epsilon}{32SH^3}$, as this bounds the fraction of trajectories in which our original bound does not apply. This is the case when $L \geq \frac{64S^2H^3W_1B}{\epsilon}$. It follows that the total variation distance between $q_{\hat{V}}$ and the empirical visitation distribution increases by at most $\frac{\epsilon}{16SH^3}$ for each step in $[H]$. If the total variation distance with $q_{\hat{V}}(\cdot, h)$ is at most $\frac{\epsilon h}{16SH^3}$ at each step, then any state with $q_{\hat{V}}(x, h) \cdot L$ expected visits gets at least $\left(q_{\hat{V}}(x, h) - \frac{\epsilon h}{16SH^3}\right) \cdot L$ visits.

Each state with $q_{\hat{V}}(x, h) \geq \frac{\epsilon}{8H^2}$ is therefore visited at least $\frac{\epsilon}{16SH^2} \cdot L$ times when the above events hold. States are locked after $\frac{16H^2WB}{\epsilon}$ visits; as such, if $L \geq \max\left(\frac{64S^2H^3W_1B}{\epsilon}, \frac{256SH^4WB}{\epsilon^2}\right)$ all states with $q_{\hat{V}}(x, h) \geq \frac{\epsilon}{8SH^2}$ are visited enough to be locked in the epoch.

□

We now have that if a state has mass at least $\frac{\epsilon}{8SH^2}$ under $q_{\hat{V}}$, it will be visited frequently enough to be locked in the epoch corresponding to value estimates $\hat{V}$, with high probability. Contrapositively, when this holds it implies that if a state is unlocked (but not reset) after the termination of an epoch, it must have had small mass under $q_{\hat{V}}$ for that epoch.

Let $U_h = \{x \mid (x, h) \text{ is unlocked at termination}\}$. We can then use a similar inductive argument (Lemma 6) to show that unlocked states have small mass under q_D at termination. An important step here is in bounding the total variation distance with $p_{\hat{V}}(\cdot; x, h)$, which we do in Lemma 5.

Lemma 5. *Let $D_{W,\hat{V}}(x, h)$ be a set of action profiles at a pair (x, h) generated by W completed runs of $\mathcal{B}$ for all players.With probability at least $1 - \delta'$, the total variation distance between $p_{\hat{V}}(\cdot; x, h)$ and (transition distribution given $a \sim D_{W,\hat{V}}(x, h)$) is at most $\frac{\epsilon}{32SH^3}$.*

Proof. Each run of $\mathcal{B}$ generates a sequence of action profiles; for each action profile, there's some fixed probability that a state x' will be visited next. Whether or not this state is actually visited is a $[0, 1]$ random variable with some expected value. The number of realized visits to x' versus the expected number of visits given the action profile can be expressed as a martingale, and as such the expectation over profile generation and transition function resampling is equal to the expected number of visits. Note that this number of visits to x' in a run of $\mathcal{B}$ is itself a random variable with mean $B \cdot p_{\hat{V}}(x'; x, h)$. and so $\mathbb{E}[p_{D,W}(x'; x, h)] = p_{\hat{V}}(x'; x, h)$. We can then apply the same concentration analysis as in Lemma 4 to give us that the total variation distance between $p_{D,W}(\cdot; x, h)$ and $p_{\hat{V}}(\cdot; x, h)$ is at most $\frac{\epsilon}{32SH^3}$ with probability $1 - \delta'$.

□

We now have that Lemma 4 and Lemma 5 hold for every window across all epochs with probability at least $1 - 2\delta/3$ by a union bound. The union-bound analysis for when Lemma 5 holds for all epochs and pairs is equivalent to that for Lemma 4.

Lemma 6. *When the algorithm terminates, with probability at least $1 - \frac{2\delta}{3}$, for each step h $\sum_{x \in U_h} q_D(x, h) \leq \frac{\epsilon h}{4H}$.*

Proof. When all events for events for Lemma 4 occur for all epochs (at most $(S+1)^H + 1$), any state which is unlocked and not reset after the end of an epoch must have $q_{\hat{V}}(x, h) \leq \frac{\epsilon}{8SH^2}$ for the corresponding $q_{\hat{V}}$. For the final epoch and its set of value estimates for all agents $\hat{V}$, this means that any unlocked state (x, h) has $q_{\hat{V}}(x, h) \leq \frac{\epsilon}{8SH^2}$ at termination, and so $\sum_{x \in U_h} q_{\hat{V}}(x, h) \leq \frac{\epsilon}{8H^2}$ for each h.

Immediately we have that the lemma holds for all pairs $(x, 1)$, as their probabilities are defined identically under q_D and $q_{\hat{V}}$.

From Lemma 5, we can see that for every locked state (x, h), we have that $d_{TV}(p_{\hat{V}}(\cdot; x', h), p_D(\cdot; x', h)) \leq \frac{\epsilon}{8H^2}$. Because we complete $\frac{16H^2W}{\epsilon}$ runs of $\mathcal{B}$ before locking any state, the total variation distance between $p_D(\cdot; x, h)$ and $p_{\hat{V}}(\cdot; x, h)$ is at most $\frac{\epsilon}{32SH^3} + \frac{\epsilon}{16H^2} \leq \frac{\epsilon}{8H^2}$, assuming worst-case total variation distance for the final sequence of up to BW trajectories for which our bound does not apply. Further, each unlocked state has mass at most $\frac{\epsilon}{8SH^2}$ under $q_{\hat{V}}$. We can bound the total variation distance between q_D and $q_{\hat{V}}$ at each step in terms of earlier steps as well as the distance from $p_{\hat{V}}$ for each pair's outgoing transition distribution using Lemma 3.

Expanding out, we can explicitly bound the total variation distance at each step, using the fact that the distributions are identical for $h = 1$. With $d^h_{q,V,D} =$

$d_{TV}(q_{\hat{V}}(\cdot,h), q_D(\cdot,h))$:

$$\begin{aligned} d^h_{q,\hat{V},D} &\leq \sum_{j=1}^{h-1} \sum_{x' \in X} q_{\hat{V}}(x',j) \cdot d_{TV}(p_{\hat{V}}(\cdot;x',j), p_D(\cdot;x',j)) \\ &\leq \sum_{j}^{h-1} \left(\sum_{x' \in U_j} q_{\hat{V}}(x',j) + \sum_{x' \in \mathcal{X} \setminus U_j} q_{\hat{V}}(x',j) \cdot \frac{\epsilon}{8H^2} \right) \\ &\leq \sum_{j}^{h-1} \left(\frac{\epsilon}{8H^2} + \frac{\epsilon}{8H^2} \right) \\ &= \frac{\epsilon(h-1)}{4H^2}. \end{aligned}$$

Applying this to our bound on the mass of unlocked states under $q_{\hat{V}}$ completes the proof of the lemma:

$$\begin{aligned} \sum_{x \in U_h} q_D(x,h) &= \sum_{x \in U_h} q_{\hat{V}}(x,h) + \sum_{x \in U_h} q_D(x,h) - q_{\hat{V}}(x,h) \\ &\leq \frac{\epsilon}{8H^2} + d_{TV}(q_{\hat{V}}(\cdot,h), q_D(\cdot,h)) \\ &\leq \frac{\epsilon h}{4H^2}. \end{aligned}$$

□

We conclude by bounding the regret when agents play according to D. First we analyze the regret each agent playing according to D under the assumption that all agents receive the maximal reward $H - h + 1$ for the remainder of the trajectory upon reaching a state in U_h. We show that this is small, and that it does not increase by much upon correcting for the unlocked states.

It will be convenient for us to consider regret with respect to function classes $\mathcal{F}^h_i : \mathcal{A}_i \times \mathcal{X} \times [h, \ldots, H] \to \mathcal{A}_i$, which we deem $\mathcal{F}^h$-regret. This is in $[0, H - h + 1]$ denoting the maximum possible downstream per-trajectory improvement by a swap function which only changes behavior in steps h and onwards. Because we complete at least W runs of $\mathcal{B}$ before locking each state, we can apply the guarantees of Corollary 3.1, where $B = B\left(\frac{\epsilon}{16H}, N\right)$ and $\eta = \frac{\epsilon}{16H^2}$ at each pair, which holds simultaneously for all pairs and players with probability $1 - \delta/3$ by a union bound, giving us a total failure probability of at most δ. For pairs at step H, which are equivalent to games with stochastic rewards, this gives us that

- the "local" $\mathcal{F}^H$-regret for a pair (x, H) is at most $\frac{\epsilon}{4H} + \frac{\epsilon}{32H^2}$, and
- the estimated value is within $\frac{\epsilon}{16H^2}$ of the true expected average value of running the bandit algorithm at that pair.

For steps $h < H$ these hold as well, but scaled by a factor of $H - h + 1$, under the assumption that estimates of pair values reflect the true expected value of being at that pair. The corresponding distribution over reward tensors for the implicitly represented game with stochastic rewards can be obtained by taking the product distribution over transition functions and reward tensors, then converting each transition-reward pair to a tensor by adding each players' value estimates for visited states at the next step to their utility (recall that rewards and transitions are independent). We will later account for this estimation error.

For every pair, there can be up to W runs of $\mathcal{B}$ at termination for which this bound doesn't hold, but otherwise we can average the contiguous sequences of W and apply the same bounds for value and regret. Because we complete at least $\frac{16WH^2}{\epsilon}$ runs of $\mathcal{B}$ before locking a state, even assuming maximal average regret for this subsequence, the total average regret increases by at most $\frac{\epsilon}{16H^2}$. The same error bound applies to value estimates.

We can then show that computed value estimates will not be far from the true expected downstream utility of that state when all agents play the correlated equilibrium. If we can bound the estimation error for downstream pairs at step $h + 1$, the estimation error at step h is bounded by the sum of the "local" and downstream error. We let $d(j)$ denote this bound for locked pairs $(x, H - j + 1)$:

$$\begin{aligned} d(1) &\leq \frac{\epsilon}{8H^2}, \\ d(j) &\leq \frac{\epsilon j}{8H^2} + d(j-1) \\ &= \sum_{i=1}^{j} \frac{\epsilon i}{8H^2} \\ &= \frac{j(j+1)}{2} \cdot \frac{\epsilon}{8H^2} \end{aligned}$$

We can also bound the regret of the distribution in a similar manner. Suppose each value estimate downstream from some pair (x, h) was exactly accurate, and each such downstream subgame had no regret; then the local regret (from the copy of $\mathcal{B}$) constitutes the entire subgame regret. Regret increases by at most twice the downstream error bound (recall we are assuming for now that this bound applies to locked and unlocked states), as this bounds the amount that any pair of swap functions $f, f' \in \mathcal{F}^h$ (including I) can deviate in the difference of their utilities when considering average reward from playing the game according to the specified action distributions. Finally, we add the downstream regret. As such, the following expression bounds the total regret at a pair:

$$\begin{aligned} (x,h) \text{ regret} \leq \text{local regret} + 2 \times \text{downstream error} \\ + \text{downstream regret} \end{aligned}$$

Here, all terms are defined with respect to the resulting distribution of profiles and the true distribution over rewards and transitions the game. We let $\hat{r}(j)$ denote the total regret

(under q_D, assuming maximal reward from unlocked states) at a pair at step $j = H - j + 1$ and let $\ell(j)$ denote the local regret. For each, we have that

$$\ell(j) = j\left(\frac{\epsilon}{4H} + 2 \cdot \frac{\epsilon}{16H^2}\right)$$

and so total regret is bounded by

$$\begin{aligned}
\hat{r}(j) &\leq \ell(j) + 2d(j-1) + \hat{r}(j-1) \\
&= \ell(j) + \sum_{i=1}^{j-1} \ell(i) + 2d(i) \\
&\leq j\left(\frac{\epsilon}{4H} + \frac{\epsilon}{8H^2}\right) + \sum_{i=1}^{j-1} i\left(\frac{\epsilon}{4H} + \frac{\epsilon}{8H^2} + \frac{(i+1)\,\epsilon}{8H^2}\right) \\
&\leq (j + j^2/2)\left(\frac{\epsilon}{4H} + \frac{\epsilon}{8H^2}\right) + \sum_{i=1}^{j-1} \frac{(i^2+i)\,\epsilon}{8H^2} \\
&\leq (j + j^2/2)\left(\frac{\epsilon}{4H} + \frac{\epsilon}{8H^2}\right) + \frac{j^3\epsilon}{24H^2}.
\end{aligned}$$

For $j = H$, corresponding to the regret bound for each state at step 1, we have that

$$\begin{aligned}
\hat{r}(H) &\leq (H + H^2/2)\left(\frac{\epsilon}{4H} + \frac{\epsilon}{8H^2}\right) + \frac{\epsilon H}{24} \\
&\leq \frac{\epsilon}{4} + \frac{\epsilon}{16} + \frac{\epsilon H}{8} + \frac{\epsilon}{8H} + \frac{\epsilon H}{24}.
\end{aligned}$$

All of the (maximal) value estimates for unlocked states are overestimates; because no swap function can improve average expected utility by more than the above bound before correcting for unlocked states, we can use the frequency of unlocked states to bound the true regret. If all unlocked states at step h have $\sum_{x \in U_h} q_D(x,h) \leq \frac{h\epsilon}{4H^2}$, their contribution to the average regret of D is bounded by

$$\begin{aligned}
\sum_{h=1}^{H}(H-h+1)\frac{h\,\epsilon}{4H^2} &= \frac{H(H+1)(H+2)}{6} \cdot \frac{\epsilon}{4H^2} \\
&\leq \frac{\epsilon H}{4}.
\end{aligned}$$

Adding in the maximal contributions from unlocked states, we have that

$$\begin{aligned}
r(H) &\leq \frac{\epsilon}{4} + \frac{\epsilon}{16} + \frac{\epsilon H}{8} + \frac{\epsilon}{8H} + \frac{\epsilon H}{24} + \frac{\epsilon H}{4} \\
&\leq 0.855\epsilon H
\end{aligned}$$

for all $\epsilon \leq 1$ and $H \geq 1$. As this bound holds simultaneously at each pair $(x, 1)$ for all players, and captures the expected regret over an entire trajectory when players play according to D, the average $\mathcal{F}$-regret *per step of the game* is less than ϵ. Thus, the policy distribution constitutes an ϵ-EFCE for the game.

□

A.5 ANALYSIS FOR FAST PLL

We restate the description of FastPLL with L given precisely.

Algorithm 4: Fast PLL. Let $B = B\left(\frac{\epsilon}{8H}, N\right)$, and the epoch length (in trajectories) be given by

$$\begin{aligned}
L \geq \frac{\sqrt{\log(2SH/\delta) + \frac{1024BH^4\gamma\log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}}}{4\gamma^2} \\
+ \frac{\log(2SH/\delta) + \frac{512BH^4\gamma\log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}}{4\gamma^2}.
\end{aligned}$$

Run H epochs, one corresponding to each step (beginning with step H) as follows:

- *Epoch for Step h:* Use a copy of $\mathcal{B}$ to select actions at each pair (x, h), augmenting rewards with computed values for pairs $(x', h+1)$ transitioned to for the next step (if $h < H$). At the end of the epoch, let $\hat{V}_i^{\mathcal{B}}(x,h)$ be the average reward received from all completed runs of $\mathcal{B}$.
- *Upstream ($h' < h$):* Select actions uniformly at random for each pair.
- *Downstream ($h' > h$):* Use $\mathcal{B}$ at each signal as in the epoch for step h', augmenting rewards with value estimates for pairs transitioned to. Restart $\mathcal{B}$ after every B rounds in which it is used, which can include rounds from a prior epoch.

Restatement of Theorem 6. *After Algorithm 3 terminates, for each pair (x,h), consider the uniform distribution over action profiles $D(x,h)$ played since epoch $H-h+1$ began. Let D be the distribution over policy profiles where the action profile for each pair (x,h) is sampled independently from $D(x,h)$. With probability at least $1-\delta$, D is an ϵ-correlated equilibrium for the game.*

Lemma 7. *With probability at least $1-\delta/2$, every state x is visited at step h at least $\frac{128BH^4\log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}$ times in epoch $H-h+1$.*

Proof. Fix some pair (x,h). Let $X = \sum_{i=1}^{L} X_i$ be a sum of indicator random variables denoting the number of times (x,h) is visited in epoch $H-h+1$. By the fast-mixing assumption, $\mathbb{E}[X] \geq \gamma L$. For L as specified, we have that

$$\gamma L - \sqrt{\frac{L\log(2SH)}{2}} \geq \frac{128BH^4\log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}$$

by the quadratic formula. By Hoeffding's inequality, with

Y being the event where $X \leq \frac{128BH^4 \log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}$:

$$\begin{aligned}
\Pr[Y] &= \Pr\left[\mathbb{E}[X] - X \geq \mathbb{E}[X] - \frac{128BH^4 \log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}\right] \\
&\leq \Pr\left[\mathbb{E}[X] - X \geq \gamma L - \frac{128BH^4 \log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}\right] \\
&\leq \Pr\left[\mathbb{E}[X] - X \geq \sqrt{\frac{L \log(2SH/\delta)}{2}}\right] \\
&\leq \exp\left(-\log(2SH/\delta)\right) \\
&\leq \frac{\delta}{2SH},
\end{aligned}$$

and the lemma follows from union-bounding over all pairs. □

Proof of Theorem 6. First we see that after epoch 1, the value estimates $\hat{V}_i(x, H)$ are within $B(\frac{\epsilon}{8H}, N)$ Let $\eta = \epsilon/8H^2$. From Lemma 7, each pair is visited at least $\frac{128BH^4 \log\left(\frac{10SHM}{\delta}\right)}{\epsilon^2}$ times in its corresponding epoch with probability at least $1 - \delta/2$. When this holds, we can apply the guarantees of Corollary 3.1, where $B = B\left(\frac{\epsilon}{8H}, N\right)$ and $\eta = \frac{\epsilon}{8H^2}$ at each pair, which holds simultaneously for all pairs and players with probability $1 - \delta/2$ by a union bound, giving us a total failure probability of δ. For pairs at step H, which are equivalent to games with stochastic rewards, this gives us that

- the "local" $\mathcal{F}^H$-regret for a pair (x, H) is at most $\frac{\epsilon}{2H} + \frac{\epsilon}{16H^2}$, and
- the estimated value is within $\frac{\epsilon}{8H^2}$ of the true expected average value of running the bandit algorithm at that pair.

Again for steps $h < H$ these hold as well, scaled by a factor of $H - h + 1$, under the assumption that estimates of pair values reflect the true expected value of being at that pair. We will account for this estimation error below.

Note that we can take these bounds to hold after all epochs terminate rather than simply the corresponding epoch. This is because neither the algorithm nor downstream values change for each step in future epochs once its value is computed. This ignores the sole possibly truncated run of $\mathcal{B}$ when the final epoch terminates. Assuming maximal average regret for this subsequence, the total average regret increases by at most $\frac{\epsilon^2}{128H^4 \log\left(\frac{10SHM}{\delta}\right)} \leq \frac{\epsilon^2}{128H^4}$ given the number of resets of $\mathcal{B}$ per epoch. The same error bound applies to value estimates.

We can then show that computed value estimates will not be far from the true expected downstream utility of that state when all agents play the correlated equilibrium. If we can bound the estimation error for downstream pairs at step $h + 1$, the estimation error at step h is bounded by the sum of the "local" and downstream error. We let $d(j)$ denote this bound for pairs $(x, H - j + 1)$:

$$\begin{aligned}
d(1) &\leq \frac{\epsilon}{8H^2} + \frac{\epsilon^2}{128H^4} \\
d(j) &\leq \frac{\epsilon j}{8H^2} + \frac{\epsilon^2 j}{128H^4} + d(j-1) \\
&= \sum_{i=1}^{j} \frac{\epsilon i}{8H^2} + \frac{\epsilon^2 i}{128H^4} \\
&= \frac{j(j+1)}{2} \cdot \left(\frac{\epsilon}{8H^2} + \frac{\epsilon^2}{128H^4}\right) \\
&\leq j^2 \left(\frac{\epsilon}{8H^2} + \frac{\epsilon^2}{128H^4}\right)
\end{aligned}$$

We can also bound the regret of the distribution in a similar manner. As in the proof of Theorem 5, the total regret at a pair can be bounded as:

$$\begin{aligned}
(x, h) \text{ regret} \leq \text{local regret} + 2 \times \text{downstream error} \\
+ \text{downstream regret}
\end{aligned}$$

Here, all terms are defined with respect to the resulting distribution of profiles and the true distribution over rewards and transitions for the game. We let $r(j)$ denote the total regret at a pair at step $j = H - j + 1$ and let $\ell(j)$ denote the local regret. For each, we have that

$$\ell(j) \leq j \left(\frac{\epsilon}{2H} + \frac{\epsilon}{16H^2} + \frac{\epsilon^2}{128H^4}\right)$$

and so total regret is bounded by

$$\begin{aligned}
r(j) &\leq \ell(j) + 2d(j-1) + r(j-1) \\
&= \ell(j) + \sum_{i=1}^{j-1} \ell(i) + 2d(i) \\
&\leq j \left(\frac{\epsilon}{2H} + \frac{\epsilon}{16H^2} + \frac{\epsilon^2}{128H^4}\right) \\
&\quad + \sum_{i=1}^{j-1} i \left(\frac{\epsilon}{2H} + \frac{\epsilon}{16H^2} + \frac{\epsilon\, i}{4H^2} + \frac{(2i+1)\epsilon^2}{128H^4}\right) \\
&\leq (j + j^2/2) \left(\frac{\epsilon}{2H} + \frac{\epsilon}{16H^2} + \frac{\epsilon^2}{128H^4}\right) \\
&\quad + \sum_{i=1}^{j-1} \left(\frac{\epsilon\, i^2}{4H^2} + \frac{\epsilon^2 i^2}{64H^4}\right) \\
&\leq (j + j^2/2) \left(\frac{\epsilon}{2H} + \frac{\epsilon}{16H^2} + \frac{\epsilon^2}{128H^4}\right) \\
&\quad + \frac{j^3}{3} \left(\frac{\epsilon}{4H^2} + \frac{\epsilon^2}{64H^4}\right)
\end{aligned}$$

For $j = H$, corresponding to the regret bound for each state at step 1, we have that

$$\begin{aligned} r(H) &\leq (H + H^2/2)\left(\frac{\epsilon}{2H} + \frac{\epsilon}{16H^2} + \frac{\epsilon^2}{128H^4}\right) \\ &\quad + \frac{H^3}{3}\left(\frac{\epsilon}{4H^2} + \frac{\epsilon^2}{64H^4}\right) \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{32} + \frac{\epsilon H}{4} + \frac{\epsilon H}{12} + \frac{\epsilon}{16H} \\ &\quad + \frac{\epsilon^2}{192H} + \frac{\epsilon^2}{256H^2} + \frac{\epsilon^2}{128H^3} \\ &\leq 0.945\epsilon H. \end{aligned}$$

As this bound holds simultaneously at each pair $(x, 1)$ for all players, and captures the expected regret over an entire trajectory, the average $\mathcal{F}^1$-regret (equivalent to $\mathcal{F}$-regret) per step of the game is less than ϵ. Thus, the policy distribution constitutes an ϵ-correlated equilibrium for the game. □

A.6 ANALYSIS FOR SINGLE-CONTROLLER STOCHASTIC GAMES

Restatement of Theorem 7. *With probability at least $1-\delta$, the uniform distribution over the sequence of policy profiles played by Algorithm 4 is an ϵ-NFCCE for the game.*

Proof of Theorem 7. The theorem follows directly from Lemma 8 and Lemma 9. □

Lemma 8. *After $T \geq \frac{8B_L(\epsilon/8)\log(M/\delta)}{\epsilon^2}$ trajectories, the controller has average regret ϵH per trajectory with probability at least $1 - \delta/M$.*

Proof of Lemma 8. Consider the sampled reward tensors (for every pair) in each trajectory. When all followers select policies in this trajectory, the current task for the controller is equivalent to an MDP (consider the fixed distribution of transitions for each action, identical across trajectories, defined by p). The task for the controller is equivalent to that of optimizing over MDPs with unknown but fixed transitions and adversarial losses; an expected per-trajectory regret bound of $\epsilon H/8$ for the *policy class* follows from Theorem 7.2 of Rosenberg and Mansour [2019] with the appropriate polynomial runtime (obtainable from inverting their regret bound), holding with respect to the set of tensors sampled in that round. Their state count corresponds to SH in our setting, as they assume a "loop-free" episodic MDP, which can be created from any MDP with an increase by a factor of at most H for the state space.

As we saw in the analysis of Theorem 3, we can again view the performance difference for each policy on the *realized* and *expected* sequence of sets of reward tensors as a martingale — given opponent policies, the reward received in the trajectory by any policy is a random variable. If $T \geq \frac{128(SH\log(N)+\log(16/\epsilon))}{\epsilon^2}$, then by Azuma-Hoeffding the probability that a policy's per-step reward deviates more than $\frac{\epsilon}{8}$ from expectation is at most $\frac{\epsilon}{8N^{SH}}$. As in the analysis of Theorem 3, by chaining deviation bounds and union-bounding over all N^{SH} policies, it then follows that the *expected* policy regret for the sequence of policy profiles, given the distribution of rewards and transitions at each state, is at most $\epsilon/2$. Given the runtime of Shifted Bandit U-CO-REPS, T is sufficiently large for this to hold extending the runtime as we did in Theorem 3. As such, for the policy sequence over $\mathcal{B}_L(\epsilon/8)$ the expected average per-step regret for the controller when sampling reward tensors and transition functions independently at each state is at most $\epsilon/2$.

Again, this is boosted to ϵ average regret with probability $1 - \frac{\delta}{M}$ after repeating for $\frac{8\log(M/\delta)}{\epsilon^2}$ such sequences, at which point the average regret is at most $\frac{3\epsilon}{4}$ with probability at least $1 - \frac{\delta}{M}$ by Hoeffding's inequality. If T is some arbitrary fixed (but sufficiently large) number of trajectories, there may be at most one run of length $B_L(\epsilon/8)$ which is *incomplete*, in that we cannot apply the above analysis; however, even assuming maximum regret across this sequence, the total average regret increases by at most $\frac{\epsilon^2}{8\log(M/\delta)} < \epsilon/4$, completing the proof. □

Lemma 9. *After $T \geq \frac{8B_F(\frac{\epsilon}{8})\log(M/\delta)}{\epsilon^2}$ trajectories, every follower has average swap regret across all pairs of at most ϵH per trajectory with probability at least $1 - \frac{\delta(M-1)}{M}$.*

Proof of Lemma 9. Followers run copies of Bayesian game algorithm $\mathcal{B}_S$ in parallel at each step, and the analysis largely follows from that in Appendix A.3. The key ideas are to observe that regret can be decomposed stepwise (any deviations cannot affect transitions), and that we did not explicitly need the distribution over signals to be static in a Bayesian game, so long as our notion of regret tracks this shifting distribution. The analysis of $\mathcal{B}_S$ in Theorem 2 carries through directly if we consider a sequence of distributions over states and we aim for small regret with respect to this sequence, as we can equivalently define martingales to track deviations from expectation for each swap function at each step. As such, after $B(\frac{\epsilon}{8S}, N)$ trajectories, the *local* expected per-step regret is at most $\frac{\epsilon}{2}$ at each step with respect to the distribution over states induced by opponents' policies at that step. As swap regret bounds traditional regret and followers' actions don't affect transitions, the expected average regret per trajectory is at most $\frac{\epsilon H}{2}$, holding with respect to the randomness in the game. Concentration analysis and handling truncation of a final sequence is equivalent to that in Lemma 8, and we union-bound over the $M - 1$ followers. □

A.7 ANALYSIS OF SIMULTANEOUS NO-REGRET WITH SHARED RANDOMNESS

Restatement of Theorem 8 *With respect to $\mathcal{F}$, PLL-SR has regret $\tilde{O}(T^{\frac{6}{7}})$ and FastPLL-SR has regret $\tilde{O}(T^{\frac{4}{5}})$.*

Proof of Theorem 8. Let T_{PLL} denote the maximum runtime of PLL (in *steps*), calibrated for an ϵ_1-EFCE. Our choice of $\epsilon_1 = \tilde{\Theta}\left(\sqrt[7]{\frac{N^3 S^{O(H)}}{T}}\right)$ is calibrated such that $T_{PLL} + \epsilon_1(T - T_{PLL}) = \tilde{\mathrm{O}}(T^{\frac{6}{7}})$. Each step after termination is equivalent to playing according to the equilibrium PLL generates, as we are sampling action profiles independently across timesteps using the shared randomness (we can use the same random string to select actions at non-visited states at that step for the purposes of defining a full policy sequence). Assuming a maximum per-step regret of 1 during the runtime of PLL (we can consider arbitrary "policies" for that window at pairs not visited in those trajectories, as PLL only chooses an action for visited pairs) and applying Theorem 5 to bound the regret for the remainder gives us the result for PLL-SR. The analysis for FastPLL-SR is symmetric. □

References

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016. doi: 10.3982/TE1808.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004. doi: 10.1109/TIT.2004.833339.

Francoise Forges. Five legitimate definitions of correlated equilibrium in games with incomplete information. CORE Discussion Papers 1993009, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 1993.

Jason Hartline, Vasilis Syrgkanis, and Éva Tardos. No-regret learning in bayesian games. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3061–3069, Cambridge, MA, USA, 2015. MIT Press.

Johan Håstad. Some optimal inapproximability results. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '97, page 1–10, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918886. doi: 10.1145/258533.258536. URL https://doi.org/10.1145/258533.258536.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *NeurIPS*, 2019.