COMPUTING EQUILIBRIA IN GAMES WITH STOCHASTIC ACTION SETS

Thomas Schwarz, Ryann Sim & Chun Kai Ling
School of Computing
National University of Singapore
tschwarz@comp.nus.edu.sg, {ryann.sim, chunkail}@nus.edu.sg

ABSTRACT

The study of learning in games typically assumes that each player always has access to all of their actions. However, in many practical scenarios, arbitrary restrictions induced by exogenous stochasticity might be placed on a player's action set. To model this setting, for a game $\mathcal{G}_{\text{orig}}$ with action set A_i for each player i, we introduce the corresponding Game with Stochastic Action Sets (GSAS) which is parametrized by a probability distribution over the players' set of possible action subsets $\mathcal{S}_i \subseteq 2^{|A_i|} \setminus \{\varnothing\}$. In a GSAS, players' strategies and Nash equilibria (NE) admit prohibitively large representations, thus existing algorithms for NE computation scale poorly. Under the assumption that action availabilities are independent between players, we show that NE in two-player zero-sum (2p0s) GSAS can be compactly represented by a vector of size $|A_i|$, overcoming naive exponential sized representation of equilibria. Computationally, we introduce an efficient approach based on *sleeping* internal regret minimization and show that it converges to approximate NE in 2p0s-GSAS at a rate $O(\sqrt{\log |A_i|/T})$ with appropriate choice of stepsizes, avoiding exponential blow-up of game-dependent constants.

1 INTRODUCTION

A common assumption in game theory is that the players always have access to all of their actions. However, in many multiagent systems, a player's actions might be randomly restricted at certain times. For instance, a trader might only have access to a subset of options on any given day due to exogenous constraints outside of their control. A naval ship might not be able to access a section of its typical patrolling area due to the presence of civilians or inclement weather. In these scenarios, a player has access to only a subset of their action space and must choose among the available actions without necessarily having full knowledge of how the availability is generated. Even in the classical setting of two-player zero-sum games, little is known about the properties of equilibrium existence and computation under stochastic action availabilities.

To model this phenomenon, we introduce the class of Games with Stochastic Action Sets (GSAS). Given a 'primal' normal-form game $\mathcal{G}_{\text{orig}}$ with action sets A_i for each player i, a GSAS is parametrized by a probability distribution $\rho \in \Delta(\mathcal{S})$. Here, $\mathcal{S} := \bigtimes_i \mathcal{S}_i$ where each $\mathcal{S}_i$ is the set of all possible action subsets $\mathcal{S}_i \subseteq 2^{|A_i|} \setminus \{\varnothing\}$ for each player i. At each timestep, Nature draws a joint availability distribution $S \in \mathcal{S}$ from ρ. Each player observes only their own available set $S_i \subseteq A_i$ and is restricted to selecting an action $a_i \in S_i$. Standard strategy formalisms in game theory require that players have to select, for each possible action subset, an action or distribution over actions to play, conditioned on the action subset observed. Clearly, this naïve representation could be prohibitively large, which motivates a study into conditions under which strategies can be compactly represented.

Beyond strategy representation, we also seek efficient methods to *compute* equilibria in GSAS. A canonical solution concept in game theory is the Nash equilibrium (NE), where no player has incentive to unilaterally deviate. However, in GSAS the problem of strategy representation renders even representing a NE in the expanded form infeasible. Computationally, a canonical connection between no-regret learning and game theoretic equilibria has led to the design of decentralized algorithms that can efficiently compute NE in two-player zero-sum games (Freund & Schapire, 1999;

Rakhlin & Sridharan, 2013; Daskalakis et al., 2011; Syrgkanis et al., 2015; Daskalakis et al., 2021). However, these algorithms typically accrue regret which scales logarithmically with the number of 'actions' in the game, which is exponential in a GSAS.

Our contributions. Motivated by the above, we study GSAS from the perspective of efficient equilibrium representation and computation. First, we show that the presence of stochastic action sets can have a significant effect on the Nash equilibria of the game. Second, we study properties of equilibria in GSAS under the assumption that player's action availabilities are independent. We show that players' strategy sets can be restricted to the set of 'implementable' strategies which contain all equilibria of the GSAS. We also establish that any implementable strategy can be compactly represented as a vector of size $|A_i|$. Third, we utilize ideas from no-regret learning in sleeping bandits to design efficient algorithms to solve GSAS. We give an algorithm called SI-MWU that minimizes a suitable notion of regret called 'Sleeping Internal' (SI)-regret in GSAS, and converges approximately to NE in 2p0s-GSAS. Using concentration arguments and a stochastic approximation procedure, we also give a method to approximately compute compact vector w that represents an approximate NE using the iterates of SI-MWU. Finally, we demonstrate the empirical efficacy of our proposed methods in GSAS. Our results demonstrate the scalability of our method compared to standard game-solvers, and exhibit approximate convergence to compact equilibria in large games.

Connections to Bayesian games. GSAS can be seen as a variation of a Bayesian game (Harsanyi, 1968), where each player has a type drawn from some distribution. Moreover, the players' available actions can depend on their types. In our setting, the available action set S_i for player i can be interpreted as their private type, drawn according to the marginal distribution induced by ρ. The player's strategy $\pi_i(S_i)$ then specifies a conditional distribution over actions given their type. However, our model is distinct in several ways from standard Bayesian games, introducing new challenges. First, classical Bayesian games assume common knowledge of payoff functions, possible types and the prior distribution over types. In contrast, we assume that ρ can be unknown to the players and that they only observe the realized subset S. This relaxes a strong assumption about common knowledge, improving the modeling power of the game class. Second, algorithms for computing equilibria in Bayesian games have rates which depend on the size on the number of possible types (Fujii, 2025; Dagan et al., 2024; Peng & Rubinstein, 2024), which is exponentially large in our setting. Hence, while GSAS share some structure with Bayesian games, they present unique challenges and motivate the development of specialized algorithms. We provide additional related work in Appendix A.

2 Games with Stochastic Action Sets

Denote the n-dimensional nonnegative quadrant by $\mathbb{R}^n_{\geq 0}$. For a finite set S, let $\Delta(S)$ be the associated probability simplex $\{x \in \mathbb{R}^{|S|}_{\geq 0} | \sum_i^{|S|} x_i = 1\}$, such that if $y \in \Delta(S), s \in S, y(s)$ is the probability that (under y) that item s is selected.

Consider an n-player normal/strategic form game $\mathcal{G}_{\text{orig}}$ with finite action set A_i, strategy profiles $A = A_1 \times \cdots \times A_n$, and utility functions $u_i : A \to [-1, 1]$. In line with prevailing conventions, we denote $a = (a_1, \ldots, a_n)$ to be a strategy profile, and as shorthand $u_i(a) = u_i(a_1, \ldots, a_n)$. We also denote by $-i$ the set of players other than i, such that $u_j(a'_i, a_{-i}) = u_j(a_1, \ldots, a'_i, \ldots, a_n)$.

Definition 2.1 (GSAS). *Given a game $\mathcal{G}_{orig} = (A_1, \ldots, A_n, u_1, \ldots, u_n)$, let $\mathcal{S}_i \subseteq 2^{|A_i|} \setminus \{\varnothing\}$ such that $\mathcal{S} = \mathcal{S}_i \times \cdots \times \mathcal{S}_n$, and $\rho \in \Delta_{\mathcal{S}}$ be a distribution over elements $\mathcal{S}$, such that $\rho(S), S \in \mathcal{S}$ gives the probability that stochastic action set S is observed. A normal form **G**ame with **S**tochastic **A**ction **S**ets is given by the tuple $\mathcal{G} = (\mathcal{G}_{orig}, \mathcal{S}, \rho)$.*

Informally, a GSAS proceeds as follows. At the start of the game, each player privately receives their action set $S_i \in \mathcal{S}_i$ from Nature based on ρ. Each then plays an action $a_i \in S_i$ simultaneously and receives a reward $u_i(a)$ based on the strategy profile $a \in A$. While each player i observes its action set S_i and might have full knowledge of ρ, they do not observe their opponents' action set S_{-i} at any point.

As described in Section 1, GSAS are a special case of Bayesian games. However, since the number of types $|\mathcal{S}_i|$ could be exponentially large in A_i, explicitly defining utilities u_i's via the Bayesian game formulation is intractable (even for $n = 2$) without exploiting the special structure afforded by

GSAS. However, defining ρ is still challenging since $\mathcal{S}$ is exponentially large in $|A_i|$ even for $n = 2$. Next, we introduce a key assumption that will benefit our analysis.

Assumption 2.2. *We assume that the $\rho(S) = \prod_i^n \rho_i(S_i)$ for some probability distributions $\rho_i : \mathcal{S}_i \to [0, 1]$, i.e., the availability of actions is independent across players.*

This assumption is often made to facilitate the analysis of the price of anarchy in Bayesian games (Fujii, 2025; Roughgarden, 2015; Syrgkanis & Tardos, 2013; Syrgkanis, 2012). Despite this assumption, a GSAS remains large since $|\mathcal{S}_i|$ remains exponential in the size of A_i.

Definition 2.3 (2p0s-GSAS)**.** *A two-player zero-sum GSAS (2p0s-GSAS) $\mathcal{G}$ is one where $\mathcal{G}_{orig}$ is two-player zero-sum, i.e., $n = 2, u_1(a) = -u_2(a)$ for all action profiles $a \in A$.*

A pure strategy in a GSAS is a deterministic mapping $\pi_i : \mathcal{S}_i \to A_i$ where $\pi_i(S_i) \in S_i$. More generally, we define a *mixed strategy* (or simply strategy) for player i as a mapping $\pi_i : \mathcal{S}_i \to \Delta(A_i)$ such that $\mathrm{supp}(\pi_i(S_i)) \subseteq S_i$. A player's strategy gives, for every possible subset of actions they could observe, a distribution of actions corresponding to the observed action subset.[1] Clearly, the representation of this strategy could be prohibitively large, and dealing with this is a key contribution.

Given a joint action set $S \in \mathcal{S}$, we denote by π the joint strategy of all players and by $\pi(a \mid S) = \prod_{i \in \mathcal{I}} \pi_i(a_i \mid S_i)$ the probability of an action profile a for every $a \in A$. The expected payoff to player i is then defined by:

$$U_i(\pi) = \mathbb{E}_{S \sim \rho} \left[\mathbb{E}_{a \sim \pi(S)}[u_i(a)] \right], \tag{1}$$

where the inner expectation is over the actions sampled independently according to each player's strategy given their available action sets, and the outer expectation is over the stochastic action set S drawn from ρ. We also define the expected payoff of a specific action a_i of player i w.r.t. the ensemble of the opponents' strategies π_{-i} by:

$$U_i(a_i; \pi_{-i}) = \mathbb{E}_{S \sim \rho} \left[\mathbb{E}_{a_{-i} \sim \pi_{-i}(S_{-i})}[u_i(a_i)] \right]. \tag{2}$$

For ease of notation, the vector of expected payoffs over all of player i's actions is denoted by $U_i(\pi_{-i})$.

Definition 2.4 (ϵ-Nash equilibrium (ϵ-NE))**.** *For $\epsilon > 0$, a strategy profile $\pi = (\pi_1, \ldots, \pi_n)$ is an ϵ-Nash equilibrium if no player can improve their expected payoff more than ϵ by unilaterally deviating from π, i.e., for all $i \in [n]$ and any of its strategies π_i':*

$$U_i(\pi_i, \pi_{-i}) \geq U_i(\pi_i', \pi_{-i}) - \epsilon, \quad \forall i \in [n],\ \forall \pi_i'. \tag{3}$$

A 0-Nash equilibrium is a Nash equilibrium, i.e., neither player can strictly benefit by unilaterally deviating.

These definitions extend the classical normal-form Nash equilibrium to GSAS and are consistent with Bayesian games and Bayesian-Nash equilibria (Harsanyi, 1968).

Remark 2.5. *Nash equilibria require that players play independently from the rest. One could also consider mediated equilibria such as (coarse) correlated equilibria in the presence of stochastic action sets. However, in this work we will focus on NE, as there are nuances in mediated equilibrium concepts in GSAS (Fujii, 2025) that lie beyond our scope.*

The presence of stochastic action availabilities for a player can have a significant effect on the set of NE, even if other players do not face stochastic action sets. We illustrate this intriguing consequence via an example:

Example 2.6. *Let $\mathcal{G}_{orig}$ be a Matching Pennies game with actions H and T. Define a GSAS by setting $u_1(H, H) = u_1(T, T) = 1$, $u_1(H, T) = u_1(T, H) = -1$, and $u_2 = -u_1$. Suppose that $\mathcal{S}_2 = \{\{H, T\}\}$, i.e., all actions are always available, but $\mathcal{S}_1 = \{S_{11} = \{H, T\}, S_{12} = \{H\}\}$ with $\rho_1(S_{11}) = \lambda$ and $\rho_1(S_{12}) = 1 - \lambda$. If $\lambda = 1$, we have the regular matching pennies game, while for $\lambda = 0$, player 1's only available action is H. The set of NE is:*

(i) $\lambda > 0.5$: player 1 sets $\pi_1(T|S_{11}) = (\lambda - 0.5)/0.5$ which "effectively" plays T with probability 0.5, while player 2 plays uniformly at random, this is essentially the standard matching pennies,

[1] Mixed strategies in Bayesian games are classically defined by *distributions over pure strategies*. It follows from classical results Harsanyi (1968) that this is strategically equivalent (for most equilibrium computation purposes, including ours) to our simpler definition of per-type conditional distribution over types.

(ii) $\lambda < 0.5$*, then player 1 sets* $\pi_1(T|S_{11}) = 1$ *and player 2 plays* T *deterministically, and*
(iii) $\lambda = 0.5$*, player 2 plays* T *with probability in* $[0.5, 1]$ *and player* 1 *selects* $\pi_1(T|S_{11}) = (\lambda - 0.5)/0.5$.

Intuitively, when $\lambda < 0.5$*, player 1 is crippled by never being able to play* T *frequently enough and player 2 takes advantage of this by playing* T *deterministically, while player* 1 *plays* T *whenever possible. In all other cases, player 1 compensates by playing* T *with higher probability when* S_{11} *is offered, i.e.,* T *is available, since they are forced to already play* H *all the time if* S_{12} *is offered.*

3 Properties of Equilibria in GSAS

For a player i, the marginal distribution over actions $a_i \in A_i$ induced by any of their strategies π_i is

$$\mathbb{P}\left[a_i; \rho_i, \pi_i\right] = \sum_{S_i \in \mathcal{S}_i} \left(\rho_i(S_i)\pi_i(a_i|S_i)\mathbb{1}\{a_i \in S_i\}\right).$$

An intuitive but crucial observation is that since action availabilities are independent and private by Assumption 2.2, player i's utility can be expressed in terms of $\mathbb{P}[a_j; \rho_j, \pi_j]$ for all players j, rather than the much larger π_j's (derivation in Appendix C.1.1):

$$U_i(\pi) = \sum_{a \in A} u_i(a) \prod_{j \in [n]} \mathbb{P}\left[a_j; \rho_j, \pi_j\right]. \tag{4}$$

Definition 3.1. *Let* $\mu_i \in \Delta(A_i)$ *be a probability distribution over player* i*'s possible actions. We say that* μ_i *is implementable if there exists a strategy* $\pi_i : \mathcal{S}_i \to \Delta(A_i)$ *such that, for every action* $a_i \in A_i$*,* $\mu_i(a_i) = \mathbb{P}[a_i; \rho_i, \pi_i]$*. In this case, we also say that* π_i *implements* μ_i*, or that* $\pi = (\pi_1, \ldots, \pi_n)$ *implements* $\mu = (\mu_1, \ldots, \mu_n)$ *if* π_i *implements* μ_i *for all* i*. The set of implementable strategies for player* i *is denoted by* $M_i \subseteq \Delta(A_i)$*.*

Given an implementable strategy μ, we abuse notation to define the expected payoff to player i following μ as: $U_i(\mu) = \mathbb{E}_{S\sim\rho}\left[\mathbb{E}_{a\sim\mu}[u_i(a)]\right]$. Similarly, we let $U_i(a_i; \mu_{-i}) = \mathbb{E}_{S\sim\rho}\left[\mathbb{E}_{a_{-i}\sim\mu_{-i}}[u_i(a_i)]\right]$. Note that by definition of μ, $\mathbb{E}_{S\sim\rho}\left[\mathbb{E}_{a\sim\pi(S)}[u_i(a)]\right] = \mathbb{E}_{S\sim\rho}\left[\mathbb{E}_{a\sim\mu}[u_i(a)]\right]$. Consequently, we can view every player's "effective" strategy space to be really over the space of M_i, implying an equivalence between Nash equilibria in the sense of Definition 2.4 and implementable marginal distributions that disincentivize unilateral deviations.

Proposition 3.2. *Consider a GSAS where* $\pi = (\pi_1, \ldots, \pi_n)$ *is a strategy profile that implements* $\mu = (\mu_1, \ldots, \mu_n)$*. Then* π *is a* ϵ*-Nash equilibrium if and only if for all* $i \in [n]$

$$U_i(\mu) \geq \max_{\mu_i' \in M_i} U_i(\mu_i', \mu_{-i}) - \epsilon.$$

Note that multiple π may implement the same μ but not vice-versa. It follows from Proposition 3.2 that if π and π' both implement μ, then if π is a NE, so is π'. In 2p0s-GSAS, an equilibrium corresponds to a bilinear saddle-point problem over M_1 and M_2, and the minimax theorem holds.

Proposition 3.3. *For a 2p0s-GSAS, a strategy profile* (π_1^*, π_2^*) *is a NE if and only if their associated* (μ_1^*, μ_2^*) *is a saddle point of the function* $U = U_1 = -U_2$*, i.e.,*

$$\mu_1^* = \operatorname*{argmax}_{\mu_1 \in M_1} \min_{\mu_2 \in M_2} U(\mu_1, \mu_2)$$

$$\mu_2^* = \operatorname*{argmin}_{\mu_2 \in M_2} \max_{\mu_1 \in M_1} U(\mu_1, \mu_2)$$

$$\textit{where } \max_{\mu_1 \in M_1} \min_{\mu_2 \in M_2} U(\mu_1, \mu_2) = \min_{\mu_2 \in M_2} \max_{\mu_1 \in M_1} U(\mu_1, \mu_2).$$

The following result shows that NE in $\mathcal{G}_{\text{orig}}$ correspond to NE in $\mathcal{G}$ if the original NE are implementable.

Proposition 3.4. *Consider GSAS* $\mathcal{G} = (\mathcal{G}_{orig}, \mathcal{S}, \rho)$*. Let* $x^* = (x_1^*, \ldots, x_n^*)$ *be a* ϵ*-NE of* $\mathcal{G}_{orig}$*, where* $x_i^* \in \Delta(A_i)$*. If* $\mu^* = x^*$ *is implementable in* $\mathcal{G}$ *by* $\pi^* = (\pi_1, \ldots, \pi_n)$*, then* π^* *is a* ϵ*-NE in* $\mathcal{G}$*.*

Note that even in the 2p0s-case, Proposition 3.4 requires the NE in $\mathcal{G}_{\text{orig}}$, x^*, to be implementable for both players, i.e., x_i^* being implementable does not imply a solution to the max-min problem (or optimal strategy for player i). See Example 2.6 (ii) for an example.

The above discussion indicates that for the purposes of equilibrium representation, we can work in the space of M_i rather than the much larger set of possible π_i. However, it is still unclear how one could recover π_i from μ_i efficiently. Our first major contribution is that every $\mu_i \in M_i$ is implemented by a compact, polynomially sized w_i.

Theorem 3.5. *Let π_i implement $\mu_i \in M_i$. Then, there exists some π' implementing μ_i and $w_i \in \Delta(A_i)$ where $\pi_i'(a_i \mid S_i) = \frac{w_i(a_i)\,\mathbb{1}\{a_i \in S_i\}}{\sum_{a_i' \in A_i} w_i(a_i')\,\mathbb{1}\{a_i' \in S_i\}}$ for all $S_i \in \mathcal{S}_i, a_i \in A_i$.*

In other words, any μ_i can be implemented by some π_i that is compactly represented by a $|A_i|-$dimensional vector w_i. It plays proportionately to w_i but restricted to the available actions S_i. Given w_i, the corresponding $\pi_i(a_i|S_i)$ for a fixed S_i can be computed in time linear in $|A_i|$.

Remark 3.6. *The compact representation w_i has two interesting properties. (i) The π_i it induces satisfies independence of irrelevant alternatives (IIA), and (ii) it yields the π_i that implements μ_i with maximum entropy. Details are deferred to Appendix C.1.6.*

Example 3.7. *Consider the 2p0s-GSAS $\mathcal{G}$ where $\mathcal{G}_{orig}$ is a variant of rock-scissors-paper where player 1 wins and loses half the amount if it plays paper, i.e., $\mathcal{G}_{orig}$ is given by:*

	***R**ock*	***S**cissors*	***P**aper*
***R**ock*	0	1	-1
***S**cissors*	-1	0	1
***P**aper*	0.5	-0.5	0

Let $\mathcal{S}_1 = \{S_{11} = \{\mathbf{R}, \mathbf{S}, \mathbf{P}\}, S_{12} = \{\mathbf{S}, \mathbf{P}\}\}$ and $\mathcal{S}_2 = \{S_{21} = \{\mathbf{R}, \mathbf{S}\}, S_{22} = \{\mathbf{S}, \mathbf{P}\}\}$ and $\rho_1(S_{11}) = \rho_1(S_{12}) = \rho_2(S_{21}) = \rho_2(S_{22}) = 0.5$. It is easy to verify that a possible NE π^* is $\pi_1^*(S_{11}) = \big(\mathbf{R} : \frac{1}{2}, \mathbf{S} : \frac{1}{2}, \mathbf{P} : 0\big)$, $\pi_1^*(S_{12}) = (\mathbf{R} : 0, \mathbf{S} : 0, \mathbf{P} : 1)$, $\pi_2^*(S_{21}) = \big(\mathbf{R} : \frac{2}{3}, \mathbf{S} : \frac{1}{3}, \mathbf{P} : 0\big)$, and $\pi_2^*(S_{22}) = \big(\mathbf{R} : 0, \mathbf{S} : \frac{1}{3}, \mathbf{P} : \frac{2}{3}\big)$. Both players obtain an expected payoff of 0. The corresponding marginal distributions of play are $\mu_1^* = \big(\mathbf{R} : \frac{1}{4}, \mathbf{S} : \frac{1}{4}, \mathbf{P} : \frac{1}{2}\big)$ and $\mu_2^* = \big(\mathbf{R} : \frac{1}{3}, \mathbf{S} : \frac{1}{3}, \mathbf{P} : \frac{1}{3}\big)$.

Utilizing Theorem 3.5 and performing some calculations gives an alternative strategy profile π' that can be represented by $w_1 = \big(\mathbf{R} : \frac{1}{2}, \mathbf{S} : \frac{1}{6}, \mathbf{P} : \frac{1}{3}\big)$, $w_2 = \big(\mathbf{R} : \frac{2}{5}, \mathbf{S} : \frac{1}{5}, \mathbf{P} : \frac{2}{5}\big)$, which represents the strategy $\pi_1(S_{11}) = \big(\mathbf{R} : \frac{1}{2}, \mathbf{S} : \frac{1}{6}, \mathbf{P} : \frac{1}{3}\big)$, $\pi_1(S_{12}) = \big(\mathbf{R} : 0, \mathbf{S} : \frac{1}{3}, \mathbf{P} : \frac{2}{3}\big)$, $\pi_2(S_{21}) = \big(\mathbf{R} : \frac{2}{3}, \mathbf{S} : \frac{1}{3}, \mathbf{P} : 0\big)$, and $\pi_2(S_{22}) = \big(\mathbf{R} : 0, \mathbf{S} : \frac{1}{3}, \mathbf{P} : \frac{2}{3}\big)$, which are consistent with the marginal distribution μ^* and also yields a NE, but this time it is compactly representable by w_1 and w_2.

4 Computing Equilibria in GSAS

In this section, we focus on the problem of NE computation in 2p0s-GSAS. There are three main regimes which are of interest. (i) The **small-support** regime, where $\mathcal{S}$ is small such that ρ is known exactly (as in Example 3.7). (ii) The **oracle-access** setting, where $\mathcal{S}$ may be exponentially large in A_i, such that ρ cannot be explicitly enumerated, but can be sampled from or queried in constant time for every $S \in \mathcal{S}$. (iii) The **sample-access** setting, where ρ is unknown and we only have access to a simulator that samples $S \sim \rho$.

In regime (i), one could in principle apply off-the-shelf solvers for Bayesian games. However, this does not take advantage of the additional structure afforded by GSAS, and typically incurs time/space costs linear in the number of types, which can be exponential in $|A_i|$. In light of this, our goal is to design a broadly applicable approach for any GSAS in regime (iii), which is the most restrictive.

4.1 No-regret learning in GSAS

For the remainder of the paper, we will focus on the *online learning* or *repeated* game paradigm. In this setting, for each player $i \in [n]$, the sequence $\{S_i^t\}_{t=1,\dots,T}$ is sampled i.i.d. following ρ_i. In each iteration t, player i observes S_i^t and plays a strategy $\pi_i^t(S_i^t)$, observing reward vector $u_i(\cdot, a_{-i}^t)$.

Notice that this setting applies to all regimes (i)-(iii) as described above. A standard performance metric in the repeated game setting is *regret*, with the folk result that no-(external)-regret algorithms (that is, algorithms that achieve sublinear regret with respect to the best fixed action in hindsight) converge in time-average to the set of Nash equilibria in 2p0s-games.

In GSAS, standard notions of regret are unsuitable since the competing action may be unavailable in certain rounds. This motivates the adaptation of modified regret definitions which have been studied in the *sleeping bandit* setting (Kleinberg et al., 2010). In this paper, we focus on *sleeping internal regret*, which was introduced by Gaillard et al. (2023).

Definition 4.1 (Sleeping Internal Regret)**.** *For any pair of actions $\hat{a}_i \in A_i$ and $\hat{a}'_i \in A_i$, the sleeping internal regret (SI-regret) for player i in T timesteps, $R^{\mathsf{INT}}_{T,i}(\hat{a}_i \to \hat{a}'_i)$, is*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{a_i^t = \hat{a}_i, \hat{a}'_i \in S_i^t\}\left(u_i(\hat{a}'_i, a^t_{-i}) - u_i(a_i^t, a^t_{-i})\right)\right]$$

where the expectation is taken over the randomness of the action availability and of the player's strategy.

In the case where a player's sleeping internal regret vanishes for each action pair, i.e. $\max_{\hat{a}_i,\hat{a}'_i} R^{\mathsf{INT}}_{T,i} = o(T)$ as $T \to \infty$, they are said to have *no-SI-regret*. The intuition is that player i does not regret not playing action $\hat{a}'_i$ (if $\hat{a}'_i$ was available) every time they played $\hat{a}_i$, for any $\hat{a}_i$, $\hat{a}'_i$. In what follows, we establish a connection between learning dynamics that achieve no-SI-regret in every action pair and the Nash equilibria of 2p0s-GSAS.

Proposition 4.2. *Consider a 2p0s-GSAS $\mathcal{G}$ where players achieve sublinear SI-regret of $R^{\mathsf{INT}}_{T,1}$ and $R^{\mathsf{INT}}_{T,2}$ after T timesteps. Define $\bar{\mu}_1 \coloneqq \frac{1}{T}\sum_{t=1}^{T} \pi_1^t(S_1^t)$ and $\bar{\mu}_2 \coloneqq \frac{1}{T}\sum_{t=1}^{T} \pi_2^t(S_2^t)$ to be the empirical marginal distributions of the players, respectively. Then, any strategy (π_1, π_2) that implements $(\bar{\mu}_1, \bar{\mu}_2)$ is a $\frac{R^{\mathsf{INT}}_1 + R^{\mathsf{INT}}_2}{T}$-approximate NE of $\mathcal{G}$.*

Remark 4.3 (On Sleeping External Regret)**.** *While sublinear external regret suffices to show time-average convergence to the set of NE in standard normal-form games, this statement does not hold in GSAS. In Appendix C.2.2, we show an example game where minimizing a natural analogue of external regret with sleeping actions known as sleeping external regret does **not** lead to a NE profile.*

Our setting lies between the bandit and full payoff feedback setting, since players have access to the payoff feedback of available actions per round. Leveraging this, we propose the Sleeping Internal Regret MWU (SI-MWU) algorithm.

Algorithm 1 SI-MWU

$E \leftarrow \{a_i \to a'_i : a_i, a'_i \in A_i, a_i \neq a'_i\}$;
$\tilde{q}^1 \leftarrow \left(\frac{1}{|E|}, \ldots, \frac{1}{|E|}\right) \in \Delta(E)$;
for $t = 1, 2, \ldots, T$ **do**
  Observe the set of available action S_i^t;
  Normalization among awake experts: $q^t(a_i \to a'_i) \leftarrow \frac{\tilde{q}^t(a_i \to a'_i)\mathbb{1}\{a'_i \in S_i^t\}}{\sum_{b_i \neq b'_i} \tilde{q}^t(b_i \to b'_i)\mathbb{1}\{b'_i \in S_i^t\}}, \forall a_i \neq a'_i$;
  Calculate $\pi_i^t(S_i^t)$ by solving the system: $\pi_i^t(S_i^t) = \sum_{a_i \neq a'_i} \pi^t_{i,a_i \to a'_i}(S_i^t) q^t(a_i \to a'_i)$;
  Play $\pi_i^t(S_i^t)$ and observe $u_i(\cdot, a^t_{-i})$;
  Update $\tilde{q}^{t+1}(a_i \to a'_i) \propto \tilde{q}^t(a_i \to a'_i) e^{(-\eta \ell^t(a_i \to a'_i))}$ {MWU with ℓ^t defined in (5)};
end for

SI-MWU is a two-level procedure outlined in Algorithm 1 where the upper level manages a vector $\pi_i^t(S_i^t) \in \Delta(A_i)$ where $\mathrm{supp}(\pi_i^t(S_i^t)) \subseteq S_i^t$. This vector is used to sample the action a_i^t at the round t. In the lower level, the algorithm maintains $|A_i|(|A_i| - 1)$ 'experts' indexed by $a_i \to a'_i$ with $a_i, a'_i \in A_i, a_i \neq a'_i$, where the expert $a_i \to a'_i$ recommends switching to a'_i whenever a_i is played. In term of expectation, this is equivalent to switching from $\pi_i^t(S_i^t)$ to a strategy $\pi^t_{i,a_i \to a'_i}(S_i^t) \in \Delta(A_i)$ where all probability mass of $\pi_i^t(S_i^t)$ on a_i is moved to a'_i. If, at the lower level, we can achieve vanishing external regret with respect to all action swaps $a_i \to a'_i$, then it implies that we

have vanishing internal sleeping regret. Hence, we use MWU at the lower level, with loss function defined as

$$\ell^t(a_i \to a_i') = \begin{cases} \hat{\ell}^t(\pi^t_{i,a_i \to a_i'}(S_i^t), a^t_{-i}), & \text{if } a_i' \in S_i^t \\ \hat{\ell}^t(\pi_i^t(S_i^t), a^t_{-i}) & \text{otherwise,} \end{cases} \tag{5}$$

where for any $p \in \Delta(A_i)$, $\hat{\ell}^t(p, a^t_{-i})$ is defined by $\hat{\ell}^t(p, a^t_{-i}) = 1 - \sum_{a_i \in A_i} p(a_i) u_i(a_i, a^t_{-i})$.

SI-MWU is closely related to algorithms in the sleeping bandits literature, and we provide additional discussion on these connections in Appendix B. In general GSAS, we show that the regret of SI-MWU taken in expectation over action availabilities and player strategies is sublinear in T.

Theorem 4.4. *For any sequence of available action sets $\{S_i^t\}_t$ and payoffs $\{u_i(\cdot, a^t_{-i})\}_t$, a player using SI-MWU with stepsizes $\eta_t = \sqrt{2 \log |A_i|}/\sqrt{t}$ enjoys SI-regret bounded by $R^{\mathsf{INT}}_{T,i}(a_i \to a_i') \leq O(\sqrt{T \log |A_i|})$ for all $a_i, a_i' \in A_i, a_i \neq a_i'$.*

Moreover, we additionally derive a probabilistic convergence statement which ensures that with high probability, the sampled regrets observed in T timesteps are close to the expected SI-regrets over all action pairs.

Proposition 4.5. *Suppose an SI-regret minimizer is run for T timesteps in a GSAS with utilities $u_i : A_i \to [-1, 1]$ and let $\tilde{R}^{\mathsf{INT}}_{T,i}(a_i \to a_i')$ denote the sampled SI-regrets for all $a_i, a_i' \in A_i, a_i \neq a_i'$. Then, for all $p \in (0, 1)$:*

$$\mathbb{P}\left[\max_{a_i, a_i'} \left| \tilde{R}^{\mathsf{INT}}_{T,i}(a_i \to a_i') - R^{\mathsf{INT}}_{T,i}(a_i \to a_i') \right| \geq \sqrt{8T \log\left(\frac{2|A_i||A_i - 1|}{p} \right)} \right] \leq p$$

Combining Theorem 4.4 with Proposition 4.5 ensures that a sampled SI-regret sequence when using SI-MWU in a general-sum GSAS has sublinear regret with high probability. Moreover, we also have that in 2p0s-GSAS, the empirical strategy distribution produced by Algorithm 1 is the marginal distribution of an ϵ-NE with high probability.

Remark 4.6. *We note that SI-MWU needs to solve a linear system at each iteration (Line 6). While this can be computationally expensive in the worst-case, in practice we observe reasonably fast per-iteration computations. In Section 5 we compare the wallclock runtimes (rather than iteration count) of SI-MWU to LP solvers, showing that SI-MWU remains scalable in large games.*

4.2 Computing compact equilibria in GSAS

In GSAS, while sublinear SI-regret guarantees convergence to an ϵ-NE π^*, the empirical distribution induced by the learning process converges to the optimal marginal distribution μ_i^*, which is not a practical representation of the NE. To deal with this, we design a procedure that can approximately compute compact vectors w_i^* associated with π^*.

Suppose that the learner/player selects a sequence of strategies $\{\pi_i^t\}_{t=1,\dots,T}$ such that $\pi_i^t(S_i^t) \to \mu_i^*$ as $T \to \infty$ where μ_i^* is a marginal distribution induced by some (approximate) Nash equilibrium π^*. Specifically, for a vector $w_i \in \mathbb{R}^{|A_i|}_{\geq 0}$, let $\hat{\pi}_i(a_i | S_i, w_i)$ be the probability of playing action a_i given availability set S_i. Then, by Theorem 3.5 there exists w_i so that $\forall S_i \in \mathcal{S}_i,\ a_i \in A_i, \hat{\pi}_i(a_i|S_i, w_i) = \frac{w_i(a_i)\mathbb{1}\{a_i \in S_i\}}{\sum_{a_i' \in A_i} w_i(a_i')\mathbb{1}\{a_i' \in S_i\}}$. Let $\hat{\mu}_i(w_i)$ be the corresponding marginal distribution where $\hat{\mu}_i(a_i|w_i) = \mathbb{E}_{S_i \sim \rho_i}[\hat{\pi}_i(a_i|S_i, w_i)]$ for all $a_i \in A_i$. The objective is to find a w_i that solves $\hat{\mu}_i(w_i) = \mu_i^*$. This can be done through a stochastic approximation as outlined in Algorithm 2, where the update is done in the log-space of w_i.

The following proposition ensures that a compact equilibrium w_i^* in 2p0s-GSAS can be computed via Algorithm 2, by using the time-averaged marginal strategies $\mu_i^t := \frac{1}{t}\sum_t \pi_i^t(S_i^t)$ instead of $\pi_i^t(S_i^t)$ at each iteration.

Proposition 4.7. *Let w_i^T be the weight vector produced by Algorithm 2. Assume that $\frac{1}{T}\sum_{t=1}^T \pi_i^t(S_i^t) \to \mu_i^*$ as $T \to \infty$, and that $\sum_{t=1}^\infty \eta_t = \infty$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$. Then, almost surely, $w_i^T \to w_i^*$ as $T \to \infty$ where w_i^* is a compact representation of a strategy that implements μ_i^*.*

Algorithm 2 Computing compact equilibrium

$\theta_i^1 \leftarrow \mathbf{1}_{|A_i|}$;
for $t = 1, 2, \ldots, T$ **do**
  Observe S_i^t and play $\pi_i^t(S_i^t)$;
  $G_i^t(a_i) \leftarrow \pi_i^t(a_i|S_i^t) - \frac{\exp(\theta_i^t(a_i)\mathbb{1}\{a_i \in S_i^t\})}{\sum_{a_i' \in S_i^t} \exp(\theta_i^t(a_i'))}$, for all $a_i \in A_i$;
  $\theta_i^{t+1} \leftarrow \theta_i^t + \eta_t G_i^t$;
end for
return $w_i^T = \exp(\theta_i^T)/\sum_{a_i \in A_i} \exp(\theta_i^T(a_i))$;

In addition to the asymptotic result above, we also utilize the *robust stochastic approximation* approach introduced by Nemirovski & Yudin (1978) and Nemirovski et al. (2009) to obtain finite-time convergence bounds. By modifying the stepsize schedule of Algorithm 2 and taking 'robust' time-averages over the iterates (details in Appendix C.4), we obtain the following high-probability result:

Proposition 4.8. *Suppose Algorithm 2 is run for T timesteps with stepsizes $1/\sqrt{t}$ on a sequence of iterates π_i^t where $\frac{1}{T}\sum_{t=1}^T \pi_i^t(S_i^t) \to \mu_i^*$ as $T \to \infty$. Let $\tilde{w}$ denote the robust time averaged value of w obtained after T timesteps. Then, for all $p \in (0,1)$, we have $\mathbb{P}\left[\|\tilde{w} - w^*\|_2^2 \geq \frac{\sqrt{2}}{p\sqrt{T}}\right] \leq p$.*

These results imply a simple procedure for computing *compact* equilibria in GSAS: for each timestep $t = 1, 2, \ldots, T$ when running SI-MWU (Algorithm 1), use the output $\pi_i^t(S_i^t)$ to update an empirical marginal distribution μ_i^t, and subsequently update a θ_i^t vector as described in Algorithm 2. In Appendix D.3, we show an example of this procedure applied to Example 3.7, recovering (approximate) w_1 and w_2 that implement a NE. Combining our results above, it follows that the procedure converges to a compact vector w_i^* associated with an approximate NE π^*.

Theorem 4.9. *Suppose SI-MWU is run for T timesteps with stepsizes $1/\sqrt{t}$ in a 2p0s-GSAS with utilities $u_i : A \to [-1, 1]$ and the empirical marginal iterates are used in Algorithm 2 to obtain compact vector $\tilde{w}_i$ using robust averaging and with stepsizes $1/\sqrt{t}$. Then, for any compact NE vector w^* and for $\delta \in (0,1)$, with probability at least $1-\delta$, $\|\tilde{w}_i - w_i^*\|_2^2 \leq O\left(\frac{1}{\delta\sqrt{T}}\right)$.*

Remark 4.10. *Our error bound for $\tilde{w}$ is given in the ℓ_2-norm. However, it is not clear how this relates to the distance from NE when using the π computed via $\tilde{w}$, i.e., $||\tilde{w}_i - w_i^*||_2^2$ could be small yet exploitable. Designing appropriate metrics that relate specifically to saddle-point residual (SPR) is left for future work. Nevertheless, we show empirically in Section 5 that our procedure computes $\tilde{w}$ vectors that compactly represent strategies with low SPR.*

Remark 4.11. *We also investigate two extensions to our results. In Appendix E, we study the use of a stochastic approximation procedure to calculate NE in* general-sum *GSAS, under the assumption that the NE of the subgames induced by the action availabilities can be solved. In Appendix F, we study an 'optimistic' modification of SI-MWU (Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015), showing that in GSAS, the naïve application of optimism does not lead to improved SI-regret bounds.*

5 Experimental Results

Experiment 1: Comparison with LP solver. We evaluate the suitability of existing game solvers for GSAS by comparing SI-MWU with the Gurobi linear program solver on the sequence form representation of a GSAS (Von Stengel, 1996). In light of Remark 4.6, we focus on comparing the wallclock time (rather than iteration count) required for SI-MWU to converge. We test on randomly generated GSAS, with payoff entries chosen uniformly at random from $[-1, 1]$. ρ is also generated randomly: each action is available uniformly at random from $[0.3, 0.5]$, and with the further restriction that at least one action must be available at each timestep (formally defined in Definition D.2).

We construct the sequence-form representation of the GSAS which encodes all possible action subsets in the game and apply Gurobi's LP solver to this expanded form. For SI-MWU, we record the wallclock time of Algorithm 1 until we reach an iterate t such that $\forall t' \in \{t, t+1, \ldots, t+1000\}, \max_{\hat{a}_i, \hat{a}_i'}[\frac{1}{t'}\tilde{R}_{t',i}^{\mathsf{INT}}(\hat{a}_i \to \hat{a}_i')] \leq 0.01$. We compare the wallclock solve time in Figure 1, and

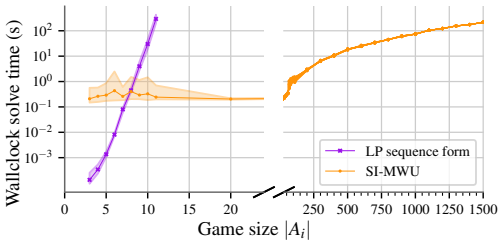


Figure 1: Wallclock time to solve randomly generated GSAS by SI-MWU and Gurobi to optimize sequence form linear program. Plot shows the average of the 20 runs with shaded region showing the range (max and min of wallclock time) of values across runs.

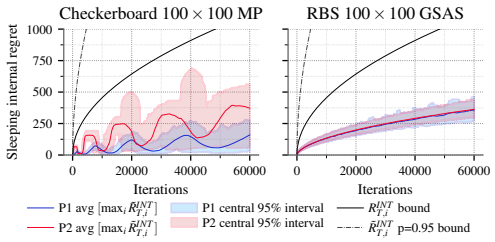
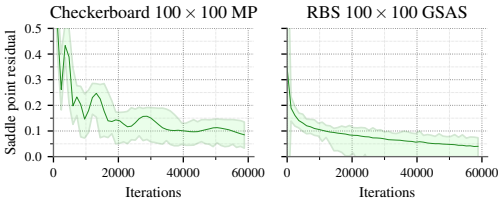
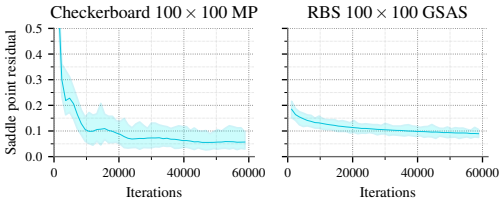


Figure 2: SI-regret from SI-MWU for several 2p0s-GSAS. For each game-type we repeat the experiment 100 times and show the average and the central 95% interval. Theoretical bounds on the expected max SI-regret and observed max SI-regret with high probability are also shown.



(a) SPR for the marginals played by SI-MWU.



(b) SPR of the Algorithm 2 computed w_i^t.

Figure 3: SPR obtained from SI-MWU for several 2p0s-GSAS. For each game we repeat the experiment 100 times and plot both the average and range (max and min regret) over the runs.

give additional experimental setup details in Appendix D.2.1. The largest random GSAS we can solve using Gurobi, of size 11×11, produces a linear program with 13313 variables and a linear constraint matrix with 126904320 nonzero entries, and took $\approx$ 293 seconds to solve on average. While SI-MWU obtained low SI-regret in much larger games, the LP solver quickly became too expensive to run, demonstrating SI-MWU's scalability.

Experiment 2: Convergence in large GSAS. In this experiment, we evaluate how SI-MWU and Algorithm 2 perform on several GSAS, presenting results for two GSAS here and on addiitional games in Appendix D.5. For all games, we repeat the experiment 100 times, sampling a new game-instance each time. Additional experimental setup details, including the sensitivity of the algorithms' performance to parameter selection, are given in Appendix D.2.2.

'RBS' is similar to a Random GSAS but with two actions available with higher probability, and a matching pennies payoff for those two actions. 'Checkerboard MP' has n actions labeled $1, 2, ..., n$ for both players and gives a payoff of 1 if both players choose actions s.t. $a_1 = a_2 \pmod 2$, and -1 if not. Player 1 always has access to all even actions, and access to one odd action chosen uniformly at random, while player 2 always has access to all actions. These games were explicitly designed to be challenging due to the payoff structure and stochastic action sets. We remark that further experiments on random GSAS (Definition D.2) exhibit lower empirical regret.

In Figure 2 we compare the observed maximum SI-regret with the theoretical bounds from Theorem 4.4 and Proposition 4.5, recording for both players at all timesteps t the observed $\max_{a_i, a_i'} \tilde{R}^{\mathsf{INT}}_{t,i}(\hat{a}_i \to \hat{a'}_i)$. Our plots show that SI-MWU achieves sublinear regret with high probability. We also show the saddle-point residual (SPR) obtained by the time-average marginals played by SI-MWU in Figure 3a, and of the strategies w_i^t obtained by Algorithm 2 run in tandem with SI-MWU in Figure 3b. Additional details on the saddle point residual, including estimating the SPR of a GSAS, are in Appendix D.2.3. We observe in our experiments that compact representations of approximate NE can still be efficiently obtained despite the probabilistic error in the SI-MWU iterates, however obtaining theoretical bounds on the SPR of w remains open.

6 Discussion and Future Work

In this paper we have taken the first step towards characterizing and computing Nash equilibria in games with stochastic action sets. A key limitation of our work is that the assumption requiring independence of action availabilities (though standard in the literature of Bayesian games) limits the applicability of our theoretical results to more general settings with coupled/dependent action distributions. Going forward, quantifying the representation error in settings with dependent action distributions, even experimentally, is an important direction for future work.

Our theoretical analysis also leaves open several interesting future research directions. These include (i) going beyond NE computation and studying notions of *correlated* equilibria in (multi-player) general-sum games, (ii) applying bandit-inspired optimism to achieve faster convergence, (iii) analyzing extensive-form and Markov games where each state has stochastic action sets, and (iv) combining sleeping-regret with practically efficient and unparametrized algorithms such as regret matching to improve empirical performance.

Acknowledgements

This project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2025) and by the National University of Singapore, under the Start-Up Grant Scheme. The authors thank Cuong Le for his contributions during the initial stage of the project.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning and game theory. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Andrea Celli, and Tuomas Sandholm. Faster no-regret learning dynamics for extensive-form correlated and coarse correlated equilibria. *arXiv preprint arXiv:2202.05446*, 2022a.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning*, pp. 536–581. PMLR, 2022b.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Michel Benaïm and Olivier Raimond. A class of self-interacting processes with applications to games and reinforced random walks. *SIAM Journal on Control and Optimization*, 48(7):4707–4730, 2010.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

Mario Bravo and Mathieu Faure. Reinforcement learning with restrictions on the action set. *SIAM Journal on Control and Optimization*, 53(1):287–312, 2015.

Andrea Celli, Alberto Marchesi, Gabriele Farina, and Nicola Gatti. No-regret learning dynamics for extensive-form correlated equilibrium. *Advances in Neural Information Processing Systems*, 33: 7722–7732, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. From external to swap regret 2.0: An efficient reduction for large action spaces. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 1216–1222, 2024.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pp. 235–254. SIAM, 2011.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34:27604–27616, 2021.

Jing Dong, Jingyu Wu, Siwei Wang, Baoxiang Wang, and Wei Chen. Taming the exponential action set: Sublinear regret and fast convergence to nash equilibrium in online congestion games. *arXiv preprint arXiv:2306.13673*, 2023.

Simone Drago, Marco Mussi, Alberto Maria Metelli, et al. Sleeping reinforcement learning. In *42nd International Conference on Machine Learning, ICML 2025*, pp. 1–60, 2025.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Stochastic regret minimization in extensive-form games. In *International Conference on Machine Learning*, pp. 3018–3028. PMLR, 2020.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Kaito Fujii. Bayes correlated equilibria, no-regret dynamics in Bayesian games, and the price of anarchy. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 2190–2191. PMLR, 2025.

Pierre Gaillard, Aadirupa Saha, and Soham Dan. One arrow, two kills: A unified framework for achieving optimal regret guarantees in sleeping bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 7755–7773. PMLR, 2023.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

John C Harsanyi. Games with incomplete information played by "Bayesian" players part ii. Bayesian equilibrium points. *Management science*, 14(5):320–334, 1968.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Satyen Kale, Chansoo Lee, and Dávid Pál. Hardness of online sleeping combinatorial optimization problems. *Advances in Neural Information Processing Systems*, 29, 2016.

Varun Kanade and Thomas Steinke. Learning hurdles for sleeping experts. *ACM Transactions on Computation Theory (TOCT)*, 6(3):1–16, 2014.

Varun Kanade, H Brendan McMahan, and Brent Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Artificial Intelligence and Statistics*, pp. 272–279. PMLR, 2009.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80(2):245–272, 2010.

Harold Joseph Kushner and Dean S Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

John Lazarsfeld, Georgios Piliouras, Ryann Sim, and Andre Wibisono. Fast and furious symmetric learning in zero-sum games: Gradient descent as fictitious play. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, pp. 3527–3577. PMLR, 2025.

MV Menon and Hans Schneider. The spectrum of a nonlinear operator associated with a matrix. *Linear Algebra and its applications*, 2(3):321–334, 1969.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, 24, 2011.

Arkadi Nemirovski and D Yudin. On cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. In *Soviet Mathematics. Doklady*, volume 19, pp. 258–269, 1978.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Gergely Neu and Michal Valko. Online combinatorial optimization with stochastic decision sets and adversarial losses. *Advances in Neural Information Processing Systems*, 27, 2014.

Quan M Nguyen and Nishant Mehta. Near-optimal per-action regret bounds for sleeping bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835. PMLR, 2024.

Ioannis Panageas, Stratis Skoulakis, Luca Viano, Xiao Wang, and Volkan Cevher. Semi bandit dynamics in congestion games: Convergence to nash equilibrium and no-regret guarantees. In *International Conference on Machine Learning*, pp. 26904–26930. PMLR, 2023.

Binghui Peng and Aviad Rubinstein. The complexity of approximate (coarse) correlated equilibrium for incomplete information games. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4158–4184. PMLR, 2024.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Tim Roughgarden. The price of anarchy in games of incomplete information. *ACM Transactions on Economics and Computation (TEAC)*, 3(1):1–20, 2015.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Aadirupa Saha, Pierre Gaillard, and Michal Valko. Improved sleeping bandits with stochastic action sets and adversarial rewards. In *International Conference on Machine Learning*, pp. 8357–8366. PMLR, 2020.

Gilles Stoltz and Gábor Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59 (1):125–159, 2005.

Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59(1):187–208, 2007.

Vasilis Syrgkanis. Bayesian games and the smoothness framework. *arXiv preprint arXiv:1203.5155*, 2012.

Vasilis Syrgkanis and Eva Tardos. Composable and efficient mechanisms. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pp. 211–220, 2013.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

APPENDIX

This supplementary material contains an overview of additional related work in Appendix A, proofs omitted from the main paper for space considerations in Appendix C, and further experimental results and details in Appendix D. Finally, we provide the some extensions to the results of the main paper in Appendix E and Appendix F.

A ADDITIONAL RELATED WORK

Sleeping regret minimization and bandits. In the multi-armed bandit (MAB) literature, the *sleeping bandit* setting studies regret minimization when actions/arms are available stochastically or adversarially (Auer et al., 2002; Blum & Mansour, 2007; Kleinberg et al., 2010; Kanade et al., 2009; Kanade & Steinke, 2014; Saha et al., 2020; Nguyen & Mehta, 2024). Sleeping action availabilities have also been applied to online combinatorial optimization (Kale et al., 2016; Neu & Valko, 2014) and reinforcement learning (Drago et al., 2025). The closest work to ours is Gaillard et al. (2023), who introduced the notion of sleeping internal regret studied in this paper.

Games with Action Set Restrictions. Benaïm & Raimond (2010) and Bravo & Faure (2015) study an algorithm called Markovian Fictitious Play (MFP) in repeated 2-player normal-form games where the action sets are restricted at each timestep. Unlike our setting, the action restrictions are dependent on the players' previous actions, and encoded via an *exploration matrix*. MFP requires players to compute a best response at each timestep, and is shown to converge a.s. to NE in (two-player) zero-sum and potential games.

Learning in Congestion Games with Exponential Action Sets. Panageas et al. (2023) and Dong et al. (2023) studied semi-bandit learning in congestion games which admit exponentially large action sets, leading to slow convergence of standard methods. While their approach also utilizes learning over a compact set of 'facilities', their setting focuses on computing NE of the original congestion game (i.e. NE which are implementable in a sense that we introduce in Proposition 3.2). Comparatively, in GSAS, the set of NE can change drastically, and we seek to compute the NE of the game induced by the stochastic action sets, not $\mathcal{G}_{\text{orig}}$.

Regret Minimization in Games. The study of regret minimization in games with structured strategy sets has a massive literature. Here, we mention several works that are directly related. Celli et al. (2020) and Anagnostides et al. (2022a) study the efficient computation of (extensive-form) correlated equilibria, while Farina et al. (2020) studies regret minimization under stochasticity.

B RELATION TO PRIOR ALGORITHMS

To achieve no-SI-regret in the adversarial sleeping experts problem, the SI-EXP3 algorithm (Gaillard et al., 2023), itself based on EXP3 (Auer et al., 2002), can be applied directly. The algorithm relies on a canonical reduction from no-internal to external regret learning (Stoltz & Lugosi, 2005; Blum & Mansour, 2007). However, SI-EXP3 is suboptimal for our setting, as it was originally designed for bandits, where reward feedback is only obtained for the chosen action. In contrast, in repeated games (without stochastic action sets) players are typically assumed to have access to the full reward feedback vector over all actions. Our setting lies between these two extremes: players have access to only the payoff feedback of available actions at each round. Hence, by redefining the loss function used in SI-EXP3, we provide an improved algorithm for our setting. In particular, the bound in Theorem 4.4 improves upon that of SI-EXP3 in Gaillard et al. (2023) by a factor of $\sqrt{|A_i|}$.

The key differences between Algorithm 1 and standard internal-regret minimization algorithms (e.g., Stoltz & Lugosi (2005)) are the definition of the loss function in Equation (5), which depends on the action availability, and the additional normalization step (Line 5) to avoid assigning positive probability mass to experts who recommend switching to unavailable actions.

C Omitted Proofs

C.1 Proofs from Section 3

C.1.1 Derivation of Equation (4)

$$
\begin{aligned}
U_i(\pi) &= \mathbb{E}_{S\sim\rho}[\mathbb{E}_{a\sim\pi(S)}[u_i(a)]] && (6)\\
&= \sum_{S\in\mathcal{S}} \left(\prod_{i\in[n]} \rho_i(S_i)\right) \sum_{a\in S} \prod_{i\in[n]} \pi_i(a_i|S_i) u_i(a) && (7)\\
&= \sum_{a\in A} \sum_{S\in\mathcal{S}:a\in S} \left(\prod_{i\in[n]} \rho_i(S_i)\right) \prod_{i\in[n]} \pi_i(a_i|S_i) u_i(a) && (8)\\
&= \sum_{a\in A} u_i(a) \sum_{S\in\mathcal{S}:a\in S} \prod_{i\in[n]} (\rho_i(S_i)\pi_i(a_i|S_i)) && (9)\\
&= \sum_{a\in A} u_i(a) \sum_{S\in\mathcal{S}} \prod_{i\in[n]} (\rho_i(S_i)\pi_i(a_i|S_i)\mathbb{1}[a_i\in S_i]) && (10)\\
&= \sum_{a\in A} u_i(a) \prod_{i\in[n]} \sum_{S_i\in\mathcal{S}_i} (\rho_i(S_i)\pi_i(a_i|S_i)\mathbb{1}[a_i\in S_i]) && (11)\\
&= \sum_{a\in A} u_i(a) \prod_{i\in[n]} \mathbb{P}\left[a_i;\rho_i,\pi_i\right] && (12)
\end{aligned}
$$

The first two lines follow by definition, the second by independence of ρ over players (Assumption 2.2) and over actions. The third step uses the fact that $\mathcal{S}_i$ contains elements that are subsets of A_i. The rest of the steps follow by algebraic manipulation.

C.1.2 Proof of Proposition 3.2

Proposition 3.2. *Consider a GSAS where $\pi = (\pi_1, \ldots, \pi_n)$ is a strategy profile that implements $\mu = (\mu_1, \ldots, \mu_n)$. Then π is a ϵ-Nash equilibrium if and only if for all $i \in [n]$*

$$U_i(\mu) \geq \max_{\mu_i'\in M_i} U_i(\mu_i', \mu_{-i}) - \epsilon.$$

Proof. ($\Longleftarrow$) Consider a strategy π which is implemented by μ. We have that $U_i(\mu) \geq \max_{\mu_i'\in M_i} U_i(\mu_i', \mu_{-i}) - \epsilon$. Expanding the expression for expected utility of π, we get:

$$
\begin{aligned}
U_i(\pi) &= \sum_{a\in A} u_i(a) \prod_{j\in[n]} \mathbb{P}[a_j;\rho_j,\pi_j] && (13)\\
&= U_i(\mu) && (14)\\
&\geq \max_{\mu_i'\in M_i} U_i(\mu_i', \mu_{-i}) - \epsilon && (15)\\
&= \sum_{a_i\in A_i} u_i(a_i)\mathbb{P}[a_i;\rho_i,\pi_i] \cdot \sum_{a_{-i}\in A_{-i}} u_{-i}(a_{-i}) \prod_{-i} \mathbb{P}[a_{-i};\rho_{-i},\pi_{-i}] - \epsilon && (16)\\
&= \max_{\pi_i'} U_i(\pi_i', \pi_{-i}) - \epsilon && (17)
\end{aligned}
$$

where we utilize the fact that $\mu_i(a_i) = \mathbb{P}[a_i;\rho_i,\pi_i]$. The proof for the forward direction is similar. □

C.1.3 PROOF OF PROPOSITION 3.3

Proposition 3.3. *For a 2p0s-GSAS, a strategy profile (π_1^*, π_2^*) is a NE if and only if their associated (μ_1^*, μ_2^*) is a saddle point of the function $U = U_1 = -U_2$, i.e.,*

$$\mu_1^* = \operatorname*{argmax}_{\mu_1 \in M_1} \min_{\mu_2 \in M_2} U(\mu_1, \mu_2)$$
$$\mu_2^* = \operatorname*{argmin}_{\mu_2 \in M_2} \max_{\mu_1 \in M_1} U(\mu_1, \mu_2)$$
$$\textit{where } \max_{\mu_1 \in M_1} \min_{\mu_2 \in M_2} U(\mu_1, \mu_2) = \min_{\mu_2 \in M_2} \max_{\mu_1 \in M_1} U(\mu_1, \mu_2).$$

Proof. The proof follows directly from von Neumann's minimax theorem (v. Neumann, 1928) and the fact that $U(\pi) = U(\mu)$ if π implements μ. □

C.1.4 PROOF OF PROPOSITION 3.4

Proposition 3.4. *Consider GSAS $\mathcal{G} = (\mathcal{G}_{orig}, \mathcal{S}, \rho)$. Let $x^* = (x_1^*, \ldots, x_n^*)$ be a ϵ-NE of $\mathcal{G}_{orig}$, where $x_i^* \in \Delta(A_i)$. If $\mu^* = x^*$ is implementable in $\mathcal{G}$ by $\pi^* = (\pi_1, \ldots, \pi_n)$, then π^* is a ϵ-NE in $\mathcal{G}$.*

Proof. Let x^* be a NE of $\mathcal{G}_{\text{orig}}$ so that $u_i(x_i^*, x_{-i}^*) \geq u_i(x_i', x_{-i}^*)$ for any $x_i' \in \Delta(A_i)$ and for all $i \in [n]$. Suppose x^* is played in the GSAS $\mathcal{G} = (\mathcal{G}_{\text{orig}}, \mathcal{S}, \rho)$. Then, the expected utility for player i when all players use x^* is:

$$U_i(x^*) = \mathbb{E}_{S \sim \rho}[\mathbb{E}_{a^* \sim x^*}[u_i(a^*)]] \tag{18}$$
$$\geq \mathbb{E}_{S \sim \rho}[\mathbb{E}_{a_i' \sim x_i'}[\mathbb{E}_{a_{-i}^* \sim x_{-i}^*}[[u_i(a_i', a_{-i}^*)]]] \tag{19}$$
$$= U(x_i', x_{-i}^*) \tag{20}$$

for all $x_i' \in \Delta(A_i)$ and $i \in [n]$. Let $\mu^* = x^*$ be implemented by some strategy π^* in $\mathcal{G}$. Then, applying Proposition 3.2 it follows that π^* is a NE of $\mathcal{G}$. □

C.1.5 PROOF OF THEOREM 3.5

Theorem 3.5. *Let π_i implement $\mu_i \in M_i$. Then, there exists some π' implementing μ_i and $w_i \in \Delta(A_i)$ where $\pi_i'(a_i \mid S_i) = \frac{w_i(a_i)\, \mathbb{1}\{a_i \in S_i\}}{\sum_{a_i' \in A_i} w_i(a_i')\, \mathbb{1}\{a_i' \in S_i\}}$ for all $S_i \in \mathcal{S}_i, a_i \in A_i$.*

Proof. The proof will require the following linear algebraic result.

Theorem C.1 (Menon & Schneider (1969))**.** *Let $X \in \mathbb{R}_{\geq 0}^{n \times m}$ be a nonnegative matrix and $r \in \mathbb{R}_{\geq 0}^n$, $c \in \mathbb{R}_{\geq 0}^m$. Then there exist $u \in \mathbb{R}_{\geq 0}^n$ and $v \in \mathbb{R}_{\geq 0}^m$ such that $P = \operatorname{diag}(u)\, X\, \operatorname{diag}(v)$ has row sums r and column sums c iff there exists a nonnegative matrix $Y \in \mathbb{R}_{\geq 0}^{n \times m}$ with row sums r, column sums c, and $\operatorname{supp}(Y) = \operatorname{supp}(X)$. Moreover, (u, v), if it exists, is unique up to a multiplicative scalar factor.*

Consider any player $i \in [n]$. We will use Theorem C.1 to show the existence of w_i. For each available set $S_i \in \mathcal{S}_i$, let us define a conditional distribution $\tilde{\pi}_i^{(\epsilon)}(S_i)$ as

$$\tilde{\pi}_i^{(\epsilon)}(a_i \mid S_i) := \begin{cases} (1-\epsilon)\, \pi_i^*(a_i \mid S_i) + \frac{\epsilon}{|S_i|} & \text{if } a_i \in \mathcal{S}_i \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where $\epsilon \in (0, 1)$ is a positive constant. Let $\tilde{\mu}_i^{(\epsilon)}$ be the corresponding marginal distribution over A_i, i.e., $\tilde{\mu}_i^{(\epsilon)}(a_i) := \sum_{S_i \in \mathcal{S}_i} \rho_i(S_i) \tilde{\pi}_i^{(\epsilon)}(a_i \mid S_i)$. Let us define a matrix $X_i \in \{0,1\}^{|\mathcal{S}_i| \times |A_i|}$ as the indicator of action availability, i.e., $X_i(S_i, a_i) := \mathbb{1}\{a_i \in S_i\}$, and a nonnegative matrix $Y_i^{(\epsilon)} \in \mathbb{R}_{\geq 0}^{|\mathcal{S}_i| \times |A_i|}$ with $Y_i^{(\epsilon)}(S_i, a_i) := \rho_i(S_i)\, \tilde{\pi}_i^{(\epsilon)}(a_i \mid S_i)$. Then for every $\epsilon > 0$, the matrix $Y_i^{(\epsilon)}$ has row sums ρ_i, column sums $\mu_i^{(\epsilon)}$, and $\operatorname{supp}(Y_i^{(\epsilon)}) = \operatorname{supp}(X_i)$. Hence the hypothesis of Theorem C.1 holds for the triple $(X_i, \rho_i, \mu_i^{(\epsilon)})$, i.e., there exist vectors $u_i^{(\epsilon)} \in \mathbb{R}_{\geq 0}^{|\mathcal{S}_i|}$ and $v_i^{(\epsilon)} \in \mathbb{R}_{\geq 0}^{|A_i|}$ such that

$$P_i^{(\epsilon)} = \operatorname{diag}\left(u_i^{(\epsilon)}\right) X_i\, \operatorname{diag}\left(v_i^{(\epsilon)}\right)$$

has row sums ρ_i and column sums $\mu_i^{(\epsilon)}$. Moreover, the pair $(u_i^{(\epsilon)}, v_i^{(\epsilon)})$ is unique up to a constant factor $\lambda^{(\epsilon)} > 0$.

For each $\epsilon > 0$, let select the value of $\lambda^{(\epsilon)}$ such that $\sum_{a_i \in A_i} v_i^{(\epsilon)}(a_i) = 1$. As a result, the set $\{v_i^{(\epsilon)} : \epsilon > 0\}$ is contained in the compact simplex $\Delta(A_i) = \{v_i \in \mathbb{R}_{\geq 0}^{|A_i|} : \sum_{a_i \in A_i} v_i(a_i) = 1\}$. It then follows by the Bolzano-Weierstrass theorem that there exists a sequence of $\epsilon_k \to 0$ such that $\{v_i^{(\epsilon_k)}\}$ converges to some $w_i \in \mathbb{R}_{\geq 0}^{|A_i|}$ with $\sum_{a_i \in A_i} w_i(a_i) = 1$. Since $\mu_i^{(\epsilon)} \to \mu_i^*$ elementwise as $\epsilon \to 0$ and the row sums of $P_i^{(\epsilon)}$ equal ρ_i for any ϵ, the (elementwise) limit $P_i = \lim_{\epsilon \to 0} P_i^{(\epsilon)}$ exists and satisfies that its row sums equal ρ_i, its column sums equal μ_i^*, and its support $\mathrm{supp}(P_i) = \mathrm{supp}(X_i)$.

Now with the existence of the limit P_i, we can define the probability distribution $\pi_i(S_i)$ over A_i for every available set S_i as

$$\begin{aligned}\pi_i(a_i \mid S_i) &:= \frac{P_i(S_i, a_i)}{\rho_i(S_i)} \\ &= \frac{X_i(S_i, a_i) w_i(a_i)}{\sum_{a_i' \in A_i} X_i(S_i, a_i') w_i(a_i')} \\ &= \frac{w_i(a_i)\, \mathbb{1}\{a_i \in S_i\}}{\sum_{a_i' \in A_i} w_i(a_i')\, \mathbb{1}\{a_i' \in S_i\}}.\end{aligned}$$

Moreover, we have the marginal probability induced by π_i as

$$\mu_i(a_i) = \sum_{S_i \in \mathcal{S}_i} \rho_i(S_i)\, \pi_i(a_i \mid S_i) = \sum_{S_i \in \mathcal{S}_i} P(S_i, a_i) = \mu_i^*(a_i).$$

That is, π_i implements μ_i^*. By similar arguments for all other players, it follows that there exists a strategy profile $\pi = (\pi_1, \ldots, \pi_n)$ where, for every player $i \in \mathcal{I}$, π_i implements μ_i^*. Hence, by Proposition 3.4, π is a Nash equilibrium. Moreover, π_i admits a unique, compact representation w_i obtained as the limit of $v_i^{(\epsilon_k)}$ as $\epsilon \to 0$. This completes the proof. □

Illustration of proof via Example 3.7. We illustrate the proof using Example 3.7 from the main text. Utilizing ideas from the proof of Theorem 3.5 and the statement of Theorem C.1, we define a matrix $X_i \in \{0,1\}^{|\mathcal{S}_i| \times |A_i|}$ as the indicator of action availability, i.e., $X_i(S_i, a_i) := \mathbb{1}\{a_i \in S_i\}$, and a nonnegative matrix $Y_i^{(\epsilon)} \in \mathbb{R}_{\geq 0}^{|\mathcal{S}_i| \times |A_i|}$ with $Y_i^{(\epsilon)}(S_i, a_i) := \rho_i(S_i)\, \tilde{\pi}_i^{(\epsilon)}(a_i \mid S_i)$ (a conditional distribution over actions, defined in Eq. 21). Then for every $\epsilon > 0$, the matrix $Y_i^{(\epsilon)}$ has row sums ρ_i, column sums $\mu_i^{(\epsilon)}$, and $\mathrm{supp}(Y_i^{(\epsilon)}) = \mathrm{supp}(X_i)$. For the row player, the above construction gives us matrices X_1 and $Y_1^{(\epsilon)}$ given by

$$X_1 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \text{ and } Y_1^{(\epsilon)} = \begin{pmatrix} \frac{3-\epsilon}{12} & \frac{3-\epsilon}{12} & \frac{\epsilon}{6} \\ 0 & \frac{\epsilon}{4} & \frac{2-\epsilon}{4} \end{pmatrix}.$$

Theorem C.1 guarantees that there exist vectors $u_i^{(\epsilon)} \in \mathbb{R}_{\geq 0}^{|\mathcal{S}_i|}$ and $v_i^{(\epsilon)} \in \mathbb{R}_{\geq 0}^{|A_i|}$ such that $P_i^{(\epsilon)} = \mathrm{diag}\left(u_i^{(\epsilon)}\right) X_i\, \mathrm{diag}\left(v_i^{(\epsilon)}\right)$ has row sums ρ_i and column sums $\mu_i^{(\epsilon)}$. In our case, by solving the system:

$$\begin{cases} \mathrm{diag}(u_1^{(\epsilon)})\, X_1\, \mathrm{diag}(v_1^{(\epsilon)})\mathbf{1} = Y_1^{(\epsilon)}\mathbf{1} \\ \mathrm{diag}(v_1^{(\epsilon)})\, X_1^\top\, \mathrm{diag}(u_1^{(\epsilon)})\mathbf{1} = {Y_1^{(\epsilon)}}^\top \mathbf{1}, \end{cases}$$

we get

$$u_1^{(\epsilon)} = \lambda \left(1, \frac{6}{3+\epsilon}\right) \text{ and}$$

$$v_1^{(\epsilon)} = \frac{1}{12\lambda} \left(3-\epsilon, \frac{(3+2\epsilon)(3+\epsilon)}{\epsilon+9}, \frac{(6-\epsilon)(3+\epsilon)}{\epsilon+9}\right)$$

for any constant $\lambda > 0$. Now set $\lambda = \frac{1}{2}$ and let $\epsilon \to 0$, we get $v_1^{(\epsilon)} \to w_1 = \left(\frac{1}{2}, \frac{1}{6}, \frac{1}{3}\right)$. With similar calculations, we obtain $w_2 = \left(\frac{2}{5}, \frac{1}{5}, \frac{2}{5}\right)$. With w_1 and w_2, we have $\pi_1(S_{11}) = \left(\frac{1}{2}, \frac{1}{6}, \frac{1}{3}\right)$,

$\pi_1(S_{12}) = \left(0, \frac{1}{3}, \frac{2}{3}\right)$, $\pi_2(S_{21}) = \left(\frac{2}{3}, \frac{1}{3}, 0\right)$, and $\pi_2(S_{22}) = \left(0, \frac{1}{3}, \frac{2}{3}\right)$, which are consistent with the marginal distribution μ^* and also yield expected payoffs of 0 for each player. Therefore, π is also a Nash equilibrium.

C.1.6 Further properties of compact representation (Remark 3.6)

(i) Independence of irrelevant alternatives (IIA): for any $S_i, S_i' \in \mathcal{S}_i$ and any $a_i, a_i' \in S_i \cap S_i'$,

$$\frac{\pi_i(a_i \mid S_i)}{\pi_i(a_i' \mid S_i)} = \frac{w_i(a_i)}{w_i(a_i')} = \frac{\pi_i(a_i \mid S_i')}{\pi_i(a_i' \mid S_i')}. \tag{22}$$

This can be interpreted as the player being consistent in their choices no matter the subset of actions seen, an intuitive consequence of Assumption 2.2. The IIA property follows directly from the relationship between w_i and its corresponding π_i.

(ii) Maximum-entropy characterization: given a marginal distribution μ_i^* induced by some equilibrium profile π^*, let Π_i be the set of all strategies that implement μ_i^*, and let $\mathcal{P}_i$ be the set of all joint distributions $Q_i(S_i, a_i)$ of (S_i, a_i) induced by ρ_i and Π_i. Then the matrix P_i constructed in the proof of Theorem 3.5 is the unique maximizer of the Shannon entropy:

$$P_i = \arg\max_{Q_i \in \mathcal{P}_i} \left\{ H(Q_i) = -\sum_{S_i \in \mathcal{S}_i, a_i \in A_i} Q_i(S_i, a_i) \log Q_i(S_i, a_i) \right\}.$$

This follows directly from the fact that P_i is the unique solution to an entropic optimal transport problem. Consequently, the strategy π_i induced by P_i is the unique conditional distribution that maximizes the Shannon entropy among all strategies implementing μ_i^*.

C.2 Proofs from Section 4

C.2.1 Proof of Proposition 4.2

Proposition 4.2. *Consider a 2p0s-GSAS $\mathcal{G}$ where players achieve sublinear SI-regret of $R_{T,1}^{\mathsf{INT}}$ and $R_{T,2}^{\mathsf{INT}}$ after T timesteps. Define $\bar{\mu}_1 \coloneqq \frac{1}{T}\sum_{t=1}^{T} \pi_1^t(S_1^t)$ and $\bar{\mu}_2 \coloneqq \frac{1}{T}\sum_{t=1}^{T} \pi_2^t(S_2^t)$ to be the empirical marginal distributions of the players, respectively. Then, any strategy (π_1, π_2) that implements $(\bar{\mu}_1, \bar{\mu}_2)$ is a $\frac{R_1^{\mathsf{INT}} + R_2^{\mathsf{INT}}}{T}$-approximate NE of $\mathcal{G}$.*

Proof. As $\bar{\mu}_1, \bar{\mu}_2$ are the empirical marginal distributions of player strategies that achieving sublinear SI-regret we have

$$\max_{\mu_1'} U_1(\mu_1', \bar{\mu}_2) - U_1(\bar{\mu}_1, \bar{\mu}_2) \le \frac{1}{T} R_{T,1}^{\mathsf{INT}}, \qquad \max_{\mu_2'} U_2(\bar{\mu}_1, \mu_2') - U_2(\bar{\mu}_1, \bar{\mu}_2) \le \frac{1}{T} R_{T,2}^{\mathsf{INT}}$$

Moreover, letting $U \coloneqq U_1 = -U_2$ and summing the above, we have

$$\max_{\mu_1'} U(\mu_1', \bar{\mu}_2) - \min_{\mu_2'} U(\bar{\mu}_1, \mu_2') \le \frac{1}{T} R_{T,1}^{\mathsf{INT}} + \frac{1}{T} R_{T,2}^{\mathsf{INT}} \tag{23}$$

The maxmin strategy can be bounded as

$$\max_{\mu_1'} \min_{\mu_2'} U(\mu_1', \mu_2') \ge \min_{\mu_2'} U(\bar{\mu}_1, \mu_2') \tag{24}$$

$$\ge \max_{\mu_1'} U(\mu_1', \bar{\mu}_2) - \frac{1}{T}\left(R_{T,1}^{\mathsf{INT}} + R_{T,2}^{\mathsf{INT}}\right) \tag{25}$$

$$\ge \min_{\mu_2'} \max_{\mu_1'} U(\mu_1', \mu_2') - \frac{1}{T}\left(R_{T,1}^{\mathsf{INT}} + R_{T,2}^{\mathsf{INT}}\right) \tag{26}$$

$$\tag{27}$$

$$\tag{28}$$

Hence by Proposition 3.3, it follows directly that $(\bar{\mu}_1, \bar{\mu}_2)$ is a $\frac{R_1^{\mathsf{INT}} + R_2^{\mathsf{INT}}}{T}$-approximate NE of $\mathcal{G}$. In particular, since any strategy (π_1, π_2) that implements $(\bar{\mu}_1, \bar{\mu}_2)$ has $U(\pi_1, \pi_2) = U(\bar{\mu}_1, \bar{\mu}_2)$, such a (π_1, π_2) is also a $\frac{R_1^{\mathsf{INT}} + R_2^{\mathsf{INT}}}{T}$-approximate NE of $\mathcal{G}$. □

C.2.2 COUNTEREXAMPLE FOR SLEEPING EXTERNAL REGRET (REMARK 4.3)

We first define a notion of Sleeping External Regret (SE-Regret) which was introduced in Blum & Mansour (2007); Kleinberg et al. (2010).

Definition C.2 (Sleeping External Regret)**.** *For any action $a_i' \in A_i$, the sleeping external regret for player i is defined as:*

$$R_{T,i}^{\mathsf{EXT}}(a_i') \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{a_i' \in S_i^t\}(u_i(a_i', a_{-i}^t) - u_i(a_i^t, a_{-i}^t))\right]$$

where the expectation is taken over the randomness of action availabilities and player strategies.

In other words, the sleeping external regret captures the amount that player i benefits if they always swapped to action a_i' for all $t \in [T]$ where possible, regardless of the original (distribution over) action a_i^t taken. Analogous to the relationship between the non-sleeping variants of internal and external regret, no-SI-Regret implies no-SE-Regret (though the converse does not hold). This follows by definitions of SI-Regret and SE-Regret:

$$R_{T,i}^{\mathsf{EXT}}(a_i') = \sum_{a_i \in A_i} R_{T,i}^{\mathsf{INT}}(a_i \to a_i').$$

We construct a single-player game $\mathcal{G}$ where $\mathcal{G}_{orig}$ has three actions a_1, a_2, a_3. Let $\mathcal{S} = \{S_1 = \{a_1, a_2\}, S_2 = \{a_2, a_3\}\}$ with $\rho(S_1) = \rho(S_2) = 0.5$. The utility function for the single player gives $u(a_1) = 1$, $u(a_2) = 2$ and $u(a_3) = 100$. Suppose the player plays uniformly in S_1 and plays a_3 w.p. 1 in S_2. The (per-iteration) SE-Regret for each action is as follows:

$$R^{\mathsf{EXT}}(a_1) = (1 - 1.5) = -0.5$$
$$R^{\mathsf{EXT}}(a_2) = (2 - 100) + (2 - 1.5) = -97.5$$
$$R^{\mathsf{EXT}}(a_3) = 0$$

The SE-Regret for the player grows sublinearly in T but the strategy is not a NE, since a profitable deviation would be to play a_2 w.p. 1 in S_1. Hence, no-SE-Regret does not suffice to guarantee convergence to NE in GSAS. We can also see that the SI-Regret for the game above is *not* sublinear in T, since the $R^{\mathsf{INT}}(a_1 \to a_2) = (2 - 1.5) = 0.5$ at each timestep.

Moreover, in the non-sleeping case where losses are adversarial, minimizing external regret does not imply minimizing internal regret (Stoltz & Lugosi, 2007). The example above shows that the analogous statement also holds in the sleeping setting.

C.2.3 PROOF OF THEOREM 4.4

Theorem 4.4. *For any sequence of available action sets $\{S_i^t\}_t$ and payoffs $\{u_i(\cdot, a_{-i}^t)\}_t$, a player using SI-MWU with stepsizes $\eta_t = \sqrt{2\log|A_i|}/\sqrt{t}$ enjoys SI-regret bounded by $R_{T,i}^{\mathsf{INT}}(a_i \to a_i') \leq O(\sqrt{T\log|A_i|})$ for all $a_i, a_i' \in A_i, a_i \neq a_i'$.*

Proof. For each round t, denote by $\mathcal{A}_t \subseteq E$ the set of awake experts. Accordingly, for each expert $e = a_i \to a_i'$ at round t we have by Equation (5):

$$\ell^t(e) = \begin{cases} \hat{\ell}^t\big(\pi_{i,e}^t(S_i^t), a_{-i}^t\big), & e \in \mathcal{A}_t, \\ \hat{\ell}^t\big(\pi_i^t(S_i^t), a_{-i}^t\big), & e \notin \mathcal{A}_t. \end{cases}$$

For every round t the following equality holds:

$$\sum_{e \in E} \tilde{q}^t(e)\, \ell^t(e) = \hat{\ell}^t\big(\pi_i^t(S_i^t), a_{-i}^t\big). \tag{29}$$

To see this, let's split the LHS of Equation (29) into awake and asleep experts (available and unavailable actions). Let $\alpha_t \coloneqq \sum_{e \in \mathcal{A}_t} \tilde{q}^t(e)$. It follows for awake experts that $\tilde{q}^t(e) = \alpha_t q^t(e)$ and by linearity of $\hat{\ell}^t(\cdot, a_{-i}^t)$ we have:

$$\sum_{e \in \mathcal{A}_t} \tilde{q}^t(e)\, \hat{\ell}^t(\pi_{i,e}^t(S_i^t), a_{-i}^t) = \alpha_t\, \hat{\ell}^t\Big(\sum_{e \in \mathcal{A}_t} q^t(e)\pi_{i,e}^t(S_i^t), a_{-i}^t\Big) = \alpha_t\, \hat{\ell}^t(\pi_i^t(S_i^t), a_{-i}^t).$$

Similarly, the asleep experts' contribution equals $(1-\alpha_t)\,\hat{\ell}^t(\pi_i^t(S_i^t), a_{-i}^t)$. Summing gives Equation (29).

By the standard multiplicative weights update upper bound (see e.g. Cesa-Bianchi & Lugosi (2006)), we have for any fixed expert $e^* \in E$, that:

$$\sum_{t=1}^{T}\sum_{e\in E}\tilde{q}^t(e)\,\ell^t(e) - \sum_{t=1}^{T}\ell^t(e^*) \leq \frac{\ln|E|}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{e\in E}\tilde{q}^t(e)\,\ell^t(e)^2. \tag{30}$$

Combining (30) with (29), it follows that

$$\begin{aligned}
\sum_{t=1}^{T}\hat{\ell}^t(\pi_i^t(S_i^t), a_{-i}^t) - \sum_{t=1}^{T}\ell^t(e^*) &\leq \frac{\ln|E|}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{e\in E}\tilde{q}^t(e)\,\ell^t(e)^2 \\
&\leq \frac{\ln|E|}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{e\in E}\tilde{q}^t(e)\,\ell^t(e) \\
&= \frac{\ln|E|}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\hat{\ell}^t(\pi_i^t(S_i^t), a_{-i}^t) \\
&\leq \frac{\ln|E|}{\eta} + \frac{\eta T}{2}.
\end{aligned}$$

Taking expectations over the action availabilities and distributions according to the algorithm yields the definition of sleeping internal regret in the LHS of above inequality. Then, by replacing $|E| = |A_i|(|A_i|-1)$ and optimizing η as per the standard proof of the MWU upper bound, we complete the proof. □

C.3 Proof of Proposition 4.5

Proposition 4.5. *Suppose an SI-regret minimizer is run for T timesteps in a GSAS with utilities $u_i : A_i \to [-1,1]$ and let $\tilde{R}_{T,i}^{\mathsf{INT}}(a_i \to a_i')$ denote the sampled SI-regrets for all $a_i, a_i' \in A_i, a_i \neq a_i'$. Then, for all $p \in (0,1)$:*

$$\mathbb{P}\left[\max_{a_i,a_i'}\left|\tilde{R}_{T,i}^{\mathsf{INT}}(a_i \to a_i') - R_{T,i}^{\mathsf{INT}}(a_i \to a_i')\right| \geq \sqrt{8T\log\left(\frac{2|A_i||A_i-1|}{p}\right)}\right] \leq p$$

Proof. We will make use of the Azuma-Hoeffding inequality (Theorem C.3) and the fact that the sampled regrets and expected regret form a *martingale difference sequence*.

Theorem C.3 (Azuma-Hoeffding Inequality (Azuma, 1967; Hoeffding, 1963))**.** *Let $Y_1, \ldots, Y_N$ be a martingale difference sequence with $a_k \leq Y_k \leq b_k$ for each k, for suitable constants a_k, b_k. Then, for any $\tau \geq 0$:*

$$\mathbb{P}\left[\sum_{k=1}^{N} Y_k \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{k=1}^{N}(b_k-a_k)^2}\right)$$

We proceed by decomposing the definition of internal regret: $R_{T,i}^{\mathsf{INT}}(a \to a')$ from Definition 4.1 encodes the sum of regrets for each possible action replacement across T samples, and there are $|A_i|(|A_i|-1)$ such random variables. Consider an arbitrary R.V. associated with action replacement $a \to a'$, $\tilde{R}_t(a \to a')$, where $t = 1, \ldots, T$ is the total number of timesteps of the algorithm. The regret for $a \to a'$ is only defined in the subset of $\{T\}$ where a, a' are available, and 0 otherwise. Denote by $\tilde{R}_t$ the instantaneous internal regret and and $\mathbb{E}[R_t]$ the expected internal regret obtained at time t, and observe that for any strategy π_t^t, $-2 \leq \tilde{R}_t - \mathbb{E}[\tilde{R}_t] \leq 2$. Moreover, $\mathbb{E}[\tilde{R}_t - \mathbb{E}[\tilde{R}_t]] = 0$, so the sequence $\{\tilde{R}_t - \mathbb{E}[\tilde{R}_t]\}_{t=1}^{T}$ is a martingale difference sequence.

Applying Theorem C.3, we get that for every action replacement pair $a, a' \in A_i$, $a \neq a'$:

$$\mathbb{P}[\tilde{R}^{\mathsf{INT}}_{T,i}(a,a') - R^{\mathsf{INT}}_{T,i}(a,a') \geq \epsilon] = \mathbb{P}[\sum_{t=1}^{T} \tilde{R}_t(a,a') - \sum_{t=1}^{T} \mathbb{E}[\tilde{R}_t(a,a')] \geq \epsilon] \tag{31}$$

$$\leq \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^{T}(2-(-2))^2}\right) \tag{32}$$

$$= \exp\left(-\frac{\epsilon^2}{8T}\right) \tag{33}$$

The inequality $\mathbb{P}[\tilde{R}^{\mathsf{INT}}_{T,i}(a,a') - R^{\mathsf{INT}}_{T,i}(a,a') \leq -\epsilon] \leq \exp\left(-\frac{\epsilon^2}{8T}\right)$ is also true, so applying the union bound we get:

$$\mathbb{P}[|\tilde{R}^{\mathsf{INT}}_{T,i}(a,a') - R^{\mathsf{INT}}_{T,i}(a,a')| \geq \epsilon] \leq 2\exp\left(-\frac{\epsilon^2}{8T}\right) \tag{34}$$

We wish to bound the probability that the maximum error over all of these R.V.s is large, which can be done again using the union bound:

$$\mathbb{P}[\max_{a,a'} |\tilde{R}^{\mathsf{INT}}_{T,i}(a,a') - R^{\mathsf{INT}}_{T,i}(a,a')| \geq \epsilon] \leq \sum_{(a,a')} \mathbb{P}[|\tilde{R}^{\mathsf{INT}}_{T,i}(a,a') - R^{\mathsf{INT}}_{T,i}(a,a')| \geq \epsilon] \tag{35}$$

$$\leq 2|A_i||A_i - 1|\exp\left(-\frac{\epsilon^2}{8T}\right) \tag{36}$$

Finally, substituting $\epsilon = \sqrt{8T\log\left(\frac{2|A_i||A_i-1|}{p}\right)}$ yields the statement.

□

C.3.1 Proof of Proposition 4.7

Proposition 4.7. *Let w_i^T be the weight vector produced by Algorithm 2. Assume that $\frac{1}{T}\sum_{t=1}^{T}\pi_i^t(S_i^t) \to \mu_i^*$ as $T \to \infty$, and that $\sum_{t=1}^{\infty}\eta_t = \infty$ and $\sum_{t=1}^{\infty}\eta_t^2 < \infty$. Then, almost surely, $w_i^T \to w_i^*$ as $T \to \infty$ where w_i^* is a compact representation of a strategy that implements μ_i^*.*

Proof. Let us rewrite the update of θ_i^t as a stochastic approximation (SA) procedure in the sense of Robbins & Monro (1951). It is well known that the asymptotic behavior of the SA iterates can be characterized by the stability of a limiting ODE (Kushner & Clark, 2012; Borkar, 2008).

Let $\hat{\pi}_i(a_i|S_i^t,\theta_i^t) = \frac{\exp(\theta_i^t(a_i)\mathbb{1}\{a_i \in S_i^t\})}{\sum_{a_i' \in S_i^t}\exp(\theta_i^t(a_i'))}$ and write $\hat{\pi}_i(S_i^t,\theta_i^t)$ as a distribution over A_i given S_i and θ_i^t. Moreover, let $\mu_i^t(a_i)$ denote the empirical marginal distribution obtained by taking the time-average of the strategies up to time t, i.e. $\mu_i^t(a_i) = \frac{1}{t}\sum_{t=1}^{t}\pi_i^t(a_i|S_i^t)$. Note that this can be updated in an online fashion while running SI-MWU and Algorithm 2 in tandem.

In our setting, we can rewrite the update of θ_i^t as:

$$\theta_i^{t+1} = \theta_i^t + \eta_t(g(\theta_i^t) + M^{t+1})$$

where $g(\theta_i^t)$ is the mean-field given by

$$g(\theta_i^t) = \mu_i^* - \mathbb{E}_{S_i\sim\rho_i}[\hat{\pi}_i(S_i,\theta_i^t)]$$

and M^{t+1} is the martingale difference given by

$$M^{t+1} = \left(\hat{\mu}_i^t - \mu_i^*\right) + \left(\mathbb{E}_{S_i\sim\rho_i}[\hat{\pi}_i(S_i,\theta_i^t)] - \hat{\pi}_i(S_i^t,\theta_i^t)\right).$$

Thus Algorithm 2 is a stochastic approximation seeking a root of $g(\theta_i) = 0$.

By construction, we have that M^{t+1} is bounded and $\mathbb{E}[M^{t+1}|\mathcal{F}_t] = 0$ where $\mathcal{F}_t = \sigma(\theta_i^\tau, S_i^\tau, \tau \leq t)$ is the filtration. Moreover, it is easy to check that $g(\theta_i)$ is a Lipschitz function. Therefore, the iterates θ_i^t are expected to track the limiting ODE

$$\dot{\theta}_i(t) = g(\theta_i(t)), t \geq 0.$$

Since $\sum_{a_i \in A_i} G_i^t(a_i) = 0$ for every t, it follows that θ_i^t lies on the hyperplane $\sum_{a_i \in A_i} \theta_i^t(a_i) = 1$ for every t. By Theorem 3.5, we know that there exists a unique θ_i^* on this hyperplane such that $g(\theta_i^*) = \mu_i^*$. Thus, it suffices to show that θ_i^* is a globally asymptotically stable equilibrium of our limiting ODE. To this end, let us define a function $V(\theta_i)$ as the KL divergence between the target joint distribution P^* with respect to S_i and a_i and the joint distribution $P(\theta_i)$ induced by θ_i (i.e., $P(\theta_i)_{S_i,a_i} = \rho_i(S_i)\hat{\pi}_i(a_i|S_i, \theta_i)$):

$$V(\theta_i) := D_{\mathrm{KL}}(P^* \,||\, P(\theta_i)).$$

Observe that $V(\theta_i) \geq 0$ for all θ_i and $V(\theta_i) = 0$ iff $\theta_i = \theta_i^*$. Moreover, $V(\theta_i)$ is continuously differentiable in θ_i and

$$\dot{V}(\theta_i(t)) = \langle \nabla_{\theta_i} V(\theta_i), \dot{\theta}_i(t) \rangle = \langle -(\mu_i^* - \mu_i(\theta)), \mu_i^* - \mu_i(\theta) \rangle = -||\mu_i^* - \mu_i(\theta_i)||^2.$$

That is, $\dot{V}(\theta_i(t)) \leq 0$ and $\dot{V}(\theta_i(t)) = 0$ iff $\mu_i^* = \mu_i(\theta_i)$ (or $\theta_i = \theta_i^*$ by the uniqueness of θ_i^*). This implies that V is a strict Lyapunov function, and thus the limiting ODE of Algorithm 2 is globally asymptotically stable. Therefore, almost surely, θ_i^t converges to θ_i^* that solves $g(\theta_i) = 0$, which implies the convergence of w_i^T, as desired.

□

C.4 Finite Time Analysis of Stochastic Approximation

The almost surely convergence above is established using the limiting ODE method (Borkar, 2008). However, for algorithmic purposes it is also useful to obtain explicit finite convergence rates (Moulines & Bach, 2011). As applied to our setting, Algorithm 2 is an instantiation of the well-known Robbins-Monro algorithm (Robbins & Monro, 1951). While asymptotic convergence to the optimal value w^* is established in Proposition 4.7, the objective is convex but not strongly convex everywhere in the domain. As such, the finite convergence rate is sensitive to the stepsize schedule (see e.g. Section 2.1 of Nemirovski et al. (2009)).

In light of this, Nemirovski & Yudin (1978) initially proposed the use of Cesaro means to avoid non-convergence/slow convergence for Lipschitz, convex functions, a method they referred to as *robust stochastic approximation*. A simple modification to Algorithm 2 can be described as follows: For any timesteps $1 \leq i \leq j$, let $\nu^t = \frac{\eta_t}{\sum_{t=i}^j \eta_t}$. We can still utilize decreasing stepsizes η_t, though the analysis holds even with constant stepsizes. Consider the points

$$\tilde{\theta}_i^j = \sum_{t=i}^{j} \nu^t \theta^t, \tag{37}$$

then, following the analysis of Nemirovski & Yudin (1978); Nemirovski et al. (2009) we can select stepsize schedule

$$\eta_t = \frac{D}{M\sqrt{t}} \tag{38}$$

where $D := \max_\theta \|\theta - \theta^1\|_2$ and M is a positive constant such that $\mathbb{E}[\|g(\theta^t)\|_2^2] \leq M^2$. In our setting, M is $\sqrt{2}$ since it is a difference between probability distributions. D is the maximal one-step difference (in terms of ℓ_2-norm) of θ from the initial condition θ^1, which is bounded by the maximal ℓ_2-norm of G_i^1. This is just the max ℓ_2 norm of a probability distribution, leading to $D = 1$.

As a direct consequence, by setting $i = 1$, $j = T$ we get

$$\mathbb{E}[g(\tilde{\theta}_1^T) - g(\theta^*)] \leq \frac{DM}{\sqrt{T}} = O\left(\frac{1}{\sqrt{T}}\right) \tag{39}$$

The above and Lipschitz continuity of g implies that the (Cesaro) averaged iterates $\tilde{\theta}_1^T$ converge with rate $O(1/\sqrt{T})$ to the minimizer θ^* in expectation,

$$\mathbb{E}[\|\tilde{\theta}_1^T - \theta^*\|_2^2] \leq O\left(\frac{1}{\sqrt{T}}\right). \tag{40}$$

This in turn guarantees that $\tilde{w}_i := \exp(\tilde{\theta}_1^T)/\sum_{a_i \in A_i} \exp(\tilde{\theta}_1^T(a_i))$ converges with rate $O(1/\sqrt{T})$ to w^*. The latter statement is true due to the fact that the softmax function is Lipschitz continuous with constant $L \leq 1$ Gao & Pavel (2017). Hence,

$$\mathbb{E}[\|\tilde{w}_i - w^*\|_2^2] = \mathbb{E}\left[\left\|\frac{\exp(\tilde{\theta}_1^T)}{\sum_{a_i \in A_i} \exp(\tilde{\theta}_1^T(a_i))} - \frac{\exp(\theta^*)}{\sum_{a_i \in A_i} \exp(\theta^*(a_i))}\right\|_2^2\right] \leq \mathbb{E}[\|\tilde{\theta}_i - \theta^*\|_2^2] \leq O\left(\frac{1}{\sqrt{T}}\right). \tag{41}$$

While the above error bound holds in expectation, we can obtain a high-probability bound on the convergence of Algorithm 2 with averaged iterates and stepsizes $1/\sqrt{t}$ to the optimal w^*.

Proposition 4.8. *Suppose Algorithm 2 is run for T timesteps with stepsizes $1/\sqrt{t}$ on a sequence of iterates π_i^t where $\frac{1}{T}\sum_{t=1}^T \pi_i^t(S_i^t) \to \mu_i^*$ as $T \to \infty$. Let $\tilde{w}$ denote the robust time averaged value of w obtained after T timesteps. Then, for all $p \in (0,1)$, we have $\mathbb{P}\left[\|\tilde{w} - w^*\|_2^2 \geq \frac{\sqrt{2}}{p\sqrt{T}}\right] \leq p$.*

The proof follows trivially from Markov's inequality. Other forms of time-averaged stepsize schedules (e.g. Polyak-Ruppert stepsizes (Polyak & Juditsky, 1992; Ruppert, 1988)) or adaptive stepsizes (Duchi et al., 2011) could provide better finite-time convergence rates, the investigation of which we leave to future work. Finally, we can combine the probabilistic error bounds for SI-MWU and Algorithm 2 into the main theorem of this section.

Theorem 4.9. *Suppose SI-MWU is run for T timesteps with stepsizes $1/\sqrt{t}$ in a 2p0s-GSAS with utilities $u_i : A \to [-1,1]$ and the empirical marginal iterates are used in Algorithm 2 to obtain compact vector $\tilde{w}_i$ using robust averaging and with stepsizes $1/\sqrt{t}$. Then, for any compact NE vector w^* and for $\delta \in (0,1)$, with probability at least $1-\delta$, $\|\tilde{w}_i - w_i^*\|_2^2 \leq O\left(\frac{1}{\delta\sqrt{T}}\right)$.*

Proof. Let Φ denote the mapping from $\mu \to \theta$. Let $\bar{\mu}_i$ denote the marginal strategies for player i obtained by time-averaging a sequence of iterates from SI-MWU. Then, the total error of $\tilde{\theta}_i$ compared to a vector θ_i^* associated with compact NE vector w_i^* can be decomposed as:

$$\|\tilde{\theta}_i - \theta_i^*\|_2^2 = \|\tilde{\theta}_i - \Phi(\bar{\mu}_i) + \Phi(\bar{\mu}_i) - \theta_i^*\|_2^2 \tag{42}$$

$$\leq \|\tilde{\theta}_i - \Phi(\bar{\mu}_i)\|_2^2 + \|\Phi(\bar{\mu}_i) - \Phi(\mu_i^*)\|_2^2 \tag{43}$$

The first term in the RHS of the inequality is the error of Algorithm 2 using the time-averaged marginals from SI-MWU as the optimal marginal distribution. The second term is the error between the Nash marginals and the time-averaged marginals.

To obtain the high probability statement on the composite error in θ-space, let $p = \delta/2$ such that for the iterates of SI-MWU, $\mathbb{P}[\text{error} \geq \frac{2C_1}{\delta\sqrt{T}}] \leq \delta/2$ and for the iterates of Algorithm 2, $\mathbb{P}[\text{error} \geq \frac{2C_2}{\delta\sqrt{T}}] \leq \delta/2$. Note that we take a looser bound on the error for SI-MWU for simplicity. Taking the union bound gives us that the probability of errors exceeding $O(1/\delta\sqrt{T})$ for both SI-MWU and Algorithm 2 are bounded above by $\delta/2 + \delta/2 = \delta$. Hence, the probability of both errors $\leq O(1/\delta\sqrt{T})$ is at least $1-\delta$.

By Proposition 4.8, the first term of Equation (43) is $\|\tilde{\theta}_i - \Phi(\bar{\mu}_i)\|_2^2 \leq 2C_2/\delta\sqrt{T}$. Moreover, the second term is:

$$\|\Phi(\bar{\mu}_i) - \Phi(\mu_i^*)\|_2^2 \leq L \cdot \|\bar{\mu}_i - \mu_i^*\| \leq \frac{L \cdot 2 \cdot C_1}{\delta\sqrt{T}}, \tag{44}$$

where L is the Lipschitz constant of the mapping Φ. Combining the two error terms as per Equation (43) and observing further that the error in w-space is bounded above by error in θ-space gives the theorem statement as required.

□

D Additional Experiments

D.1 Game definitions

In this section we formally define all the games used in our experiments.

Definition D.1 (Checkerboard $n \times n$ matching pennies ('Checkerboard MP')). *With n even, we define Checkerboard $n \times n$ matching pennies as a 2p0s-GSAS with $|A_i| = n$ actions for each player with payoffs*

$$A_{i,j} := \begin{cases} 1 & \text{if } i = j \pmod 2 \\ -1 & \text{otherwise} \end{cases}$$

and action availabilities for player 1 drawn uniformly at random from

$$\{i \cup \{0, 2, 4, ..., n\} \mid i \in \{1, 3, ..., n-1\}\}$$

and player 2 always having access to all actions.

Intuitively, player 1 wins if they pick an action with the same value $\pmod 2$ as player 2, and player 1 has access to all even actions and a single odd action chosen uniformly at random, while player 2 has access to all actions.

Definition D.2 (Random 2p0s-GSAS). *A random GSAS has $|A_i| = n$ actions for each player, with payoffs given by*

$$A_{i,j} \sim \text{Uniform}[-1, 1]$$

and for each action we sample $p_{a_i} \sim \text{Uniform}[3/10, 5/10]$. ρ then selects actions a_i to be available independently with probability p_{a_i}, redrawing if no action is present.

Definition D.3 (Random Biased Support ('RBS') GSAS). *A RBS has $|A_i| = n$ actions for each player, with payoffs given by*

$$A_{i,j} \sim \text{Uniform}[-1, 1]$$

except for the first two actions which have a matching pennies structure given by $A_{1,1} = A_{2,2} = 1$ and $A_{1,2} = A_{2,1} = 0$. For the first two actions we sample $p_{a_i} \sim \text{Uniform}[3/10, 5/10]$ and for the others we sample $p_{a_i} \sim \text{Uniform}[1/100, 2/100]$. ρ then selects actions a_i to be available independently with probability p_{a_i}, redrawing if no action is present.

Definition D.4 (Biased $n \times n$ Rock-Paper-Scissors ('Biased RPS')). *Biased $n \times n$ RPS has $|A_i| = n$ actions for each player, with payoffs given by*

$$A_{i,j} := \begin{cases} -1 & \text{if } j = i + 1 \pmod n \\ 1 & \text{if } j = i - 1 \pmod n \\ 0 & \text{otherwise} \end{cases}$$

Player 1 only has their first action available with probability $\frac{2}{n}$, and otherwise has all their actions available. Player 2 always has all actions available.

This is a special case of the generalized $n \times n$ Rock-Paper-Scissors defined in Lazarsfeld et al. (2025).

Definition D.5 (Biased $n \times n$ Matching Pennies ('Biased MP')). *Biased $n \times n$ MP has $|A_i| = n$ actions for both players and payoffs given by*

$$A_{i,j} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Player 1 has w.p. $4/5$ only actions $\{a_1, \ldots, a_{\lfloor \frac{3n}{5} \rfloor}\}$ available, and w.p. $1/5$ only actions $\{a_{\lfloor \frac{2n}{5} \rfloor}, \ldots, n\}$ available. Player 2 always has all actions available.

D.2 Additional Experimental Details

All experiments were run on a 2021 MacBook Pro with 32 GB of RAM and an 'Apple M1 Pro' chip with 8 cores. The Gurobi commerical solver was allowed to use any number of threads. Gurobi optimizer 'version 13.0.1 build v13.0.1rc0 (mac64[arm] - Darwin 24.6.0 24G90)' was used. Python version 3.13.3 with numpy version 2.4.1 and scipy 1.17.0 were used. Central 95% intervals, where reported, were computed using 'numpy.quantile'.

Table 1: η coefficients used in Experiment 2, with SI-MWU using $\eta_t = H\sqrt{\frac{\log|A_i|(|A_i|-1)}{t}}$ and Algorithm 2 using $\eta_t = \frac{K}{\sqrt{t}}$

GAME	SI-MWU H COEFFICIENT	ALGORITHM 2 K COEFFICIENT
100×100 CHECKERBOARD MP	32	10
100×100 RBS GSAS	3	10
EXAMPLE 3.7 GAME	1	$\sqrt{2}$
100×100 BIASED RPS	12	10
100×100 BIASED MP	8	10

Table 2: Maximum SI-Regret from SI-MWU across all actions across 60,000 iterations in 100×100 Checkerboard Matching Pennies using $\eta_t = H\sqrt{\frac{\log|A_i|(|A_i|-1)}{t}}$. For each choice of H, we repeat 20 times (with a newly generated game) and show the mean, minimum and maximum across these repetitions, rounded to whole numbers for clarity. We observe any choice of $H \in [1, 50]$ achieves reasonable max SI-Regret for both players.

SI-MWU H	P1 MEAN	P1 RANGE	P2 MEAN	P2 RANGE
0.01	9	9 − 10	938	938 − 938
0.1	45	44 − 46	115	115 − 115
1	22	21 − 24	516	515 − 518
3	410	349 − 482	193	192 − 195
5	367	319 − 434	148	142 − 151
10	287	216 − 392	71	63 − 76
30	234	183 − 315	41	29 − 48
32	235	169 − 383	42	31 − 50
50	321	222 − 376	48	37 − 70
75	465	281 − 626	51	28 − 62
100	10698	134 − 106243	34	3 − 77
200	6121	5 − 119555	1435	2 − 2222

D.2.1 EXPERIMENT 1: COMPARISON WITH LP SOLVER

In Figure 1 we plot the wallclock runtime reported by Gurobi, and note that this does not include the time to construct the sequence form representation or the time to build the Gurobi model, which in practice is also slow for large games. This was done in order to demonstrate that SI-MWU scales better than Gurobi. We run SI-MWU with $\eta_t = \sqrt{\frac{\log|A_i|(|A_i|-1)}{t}}$.

The sequence form linear program has $2^n + n2^{n-1} + 1$ many variables, and a constraint matrix with that many rows and columns, which quickly becomes infeasible to solve as observed. An interesting point that we observed in our experiments on Random GSAS is that SI-MWU often obtains very low SI-regret compared to the other games shown in Section 5. Nevertheless, we observe empirically that other GSAS games still scale well when solved by SI-MWU.

D.2.2 EXPERIMENT 2: CONVERGENCE OF ALGORITHM 1 AND 2

We find using higher η values in SI-MWU and Algorithm 2 causes faster convergence in practice, and so run SI-MWU with $\eta_t = H\sqrt{\frac{\log|A_i|(|A_i|-1)}{t}}$, and Algorithm 2 with $\eta_t = \frac{K}{\sqrt{t}}$ where H and K is given for each game in Table 1. Convergence is not too sensitive to these parameters, as shown in Table 2 and Table 3 which give the impact of different coefficient choices for 100×100 Checkerboard Matching Pennies. In Algorithm 2 we also ignore the first 500 datapoints produced by SI-MWU, finding that this increases how fast we converge. This is because the first few time averaged marginals μ_i^t produced by SI-MWU can, and often do, oscillate significantly from the equilibrium marginals μ_i^*. Table 4 shows the impact on the saddle point residual of the learned compact w representation from dropping different numbers of initial datapoints.

Table 3: Saddle point residual of the compact w learned by Algorithm 2 when using different step sizes, after 60,000 iterations in 100×100 Checkerboard Matching Pennies. Algorithm 2 is run with the corresponding $\eta_t = \frac{K}{\sqrt{t}}$ for each entry, and we run SI-MWU with $H = 32$. For each choice of K, we repeat 20 times (with a newly generated game) and show the mean, minimum and maximum SPR across these repetitions.

ALG 2 K	MEAN SPR	SPR RANGE
5	0.23336	0.19159 −0.27776
10	0.15755	0.11215 −0.20654
15	0.12541	0.07864 −0.17623
20	0.10691	0.05942 −0.15876

Table 4: Saddle point residual of the compact w learned by Algorithm 2 when skipping various numbers of initial datapoints, after 60,000 iterations in 100×100 Checkerboard Matching Pennies using $K = 10$ for Algorithm 2 and $H = 32$ for SI-MWU. For each choice of skipped datapoints, we repeat 20 times (with a newly generated game) and show the mean, minimum and maximum SPR across the repetitions.

ALG 2 NUM. DATAPOINTS TO SKIP	MEAN SPR	SPR RANGE
1	0.53145	0.50588 −0.55794
50	0.06757	0.04435 −0.11937
100	0.05295	0.03685 −0.10046
500	0.0631	0.03334 −0.11381
5000	0.10695	0.05154 −0.16343

Both Checkerboard MP and RBS were chosen because they are large games where players can incur high SI-regret. Checkerboard MP in particular was explicitly designed to be hard for SI-MWU to solve: P1 is given $\frac{n}{2}$ even actions and only one odd action. If P1 regrets not choosing the odd action, all $\frac{n}{2}$ even actions will incur regret at that iteration. Furthermore, each individual's odd actions are seen infrequently and so the corresponding experts are not updated frequently in expectation.

D.2.3 CALCULATING THE SADDLE-POINT RESIDUAL IN EXPERIMENTS

Since we focus on the 2p0s case, we can consider the *saddle-point residual* of a strategy pair (π_1, π_2). In particular, let $U = U_1 = -U_2$. Then we have

$$\begin{aligned} SPR(\pi_1, \pi_2) &= [U(\pi_1, \pi_2) - \min_{\pi_2'} U(\pi_1, \pi_2')] + [\max_{\pi_1'} U(\pi_1', \pi_2) - U(\pi_1, \pi_2)] \\ &= \max_{\pi_1'} U(\pi_1', \pi_2) - \min_{\pi_2'} U(\pi_1, \pi_2') \end{aligned}$$

It is easy to see that the saddle-point residual of (π_1, π_2) is zero if and only if it is a Nash equilibrium.

We wish to calculate the saddle-point residual of strategies π_i produced by Algorithm 2 and by marginal distributions produced by Algorithm 1, however as $U(\pi_1, \pi_2) = \mathbb{E}_{S\sim\rho}\left[\mathbb{E}_{a\sim\pi(S)}[u_i(a)]\right]$ it is infeasible to calculate when we only have sampling access to ρ or when the support of ρ is too large to enumerate. We thus wish to calculate where possible, and estimate when not, the saddle point residual in three different regimes: (i) ρ is known and has small support, (ii) only sampling access to ρ is available, and (iii) when ρ is such that each action is available independently across the action set. For each regime, we describe the calculation only for $\max_{\pi_1'} U(\pi_1', \pi_2)$, noting the procedure for $\min_{\pi_2'} U(\pi_1, \pi_2')$ is similar.

(i) Known ρ with small support regime. If we know $\rho = (\rho_1, \rho_2)$ we can directly calculate $\max_{\pi_1'} U(\pi_1', \pi_2)$. For a marginal distribution μ_2 induced by π_2 we compute

$$\max_{\pi_1'} U(\pi_1', \pi_2) = \sum_{S_1 \in \rho_1} \mathbb{P}[S_1] \max_{a_1 \in S_1} \left[\sum_{a_i \in A_2} \mu_i U(a_1, a_i) \right] \tag{45}$$

If we instead have a strategy π_2, we can first calculate the expected payoff for each action $a_i \in A_1$ as

$$\mathbb{E}_{S_2 \in \rho_2}[U(a_i, \pi_2)] = \sum_{S_2 \in \rho_2} \mathbb{P}[S_2] U(a_i; \pi_2(S_2))$$

and can then proceed the same as for the marginal.

We use this procedure for calculating the SPR for Example 3.7 and for small ($n < 10$) instances of 'Checkerboard MP', 'Biased RPS', and 'Biased MP'.

(ii) Sample ρ access regime. If we only have sample access to ρ - for example when enumerating all possibly observed $S_i \in \rho_i$ is infeasible - we repeat the same process as in the known ρ regime but sampling from ρ as needed, instead of enumerating all possible action availabilities, to estimate the SPR.

We use this regime to estimate the SPR for large ($n > 10$) instances of 'Checkerboard MP', 'Biased RPS', and 'Biased MP'.

(iii) Actions available independently regime. If a players actions are available independently, and we know the probability that an action is available, after estimating the expected payoff for our opponent $E(a_i) := \mathbb{E}_{S_2 \in \rho_2}[U(a_i, \pi_2)]$ for each action $a_i \in A_1$ the same as in the *Sample ρ access regime* we can then directly calculate $\mathbb{E}_{S_1 \sim \rho_1}[\max_{\pi_1'} E(a_i)]$. Let $p_{a_i} := \mathbb{P}[a_i \in S_i]$.

Assume w.l.o.g. that a_i is sorted such that $E(a_j) \geq E(a_{j+1})$. Ignoring the case no action is present and we need to resample S_1, action a_j contributes

$$\prod_{i<j} \mathbb{P}[a_i \notin S_1 \mid S_1 \neq \emptyset] \mathbb{P}[a_j \in S_1 \mid S_1 \neq \emptyset] E(a_j)$$

to $\max_{\pi_1'} U(\pi_1', \pi_2)$. We account for the case we need to resample by solving:

$$\max_{\pi_1'} U(\pi_1', \pi_2) = \sum_{j=1}^{|A_1|} \left\{ \prod_{i<j} \mathbb{P}[a_i \notin S_1 \mid S_1 \neq \emptyset] \cdot \mathbb{P}[a_j \in S_1 \mid S_1 \neq \emptyset] E(a_j) \right\} + \mathbb{P}[S_1 = \emptyset] \left(\max_{\pi_1'} U(\pi_1', \pi_2) \right)$$

$$\max_{\pi_1'} U(\pi_1', \pi_2) = \frac{1}{1 - \prod_{a_i \in S_1} p_{a_i}} \sum_{j=1}^{|A_1|} \left\{ \prod_{i<j} \mathbb{P}[a_i \notin S_1 \mid S_1 \neq \emptyset] \mathbb{P}[a_j \in S_1 \mid S_1 \neq \emptyset] E(a_j) \right\}$$

We use this to more accurately estimate the saddle point residual of a random 2p0s-GSAS (Definition D.2).

D.3 Solving Example 3.7 Computationally

Recall the payoff matrix of the RSP game from Example 3.7:

	Rock	**S**cissors	**P**aper
Rock	0	1	-1
Scissors	-1	0	1
Paper	0.5	-0.5	0

We apply the procedure outlined in Section 4. In particular, we run the SI-MWU algorithm and at each step, update the θ^t value according to Algorithm 2. We show the learnt weights in every iterate output by running Algorithm 2 on the output of Algorithm 1 in Figure 4, with and without robust averaging. We find the marginals played by SI-MWU and the weights learned by Algorithm 2 converge to the expected values.

We also analyze Example 3.7 in Appendix D.5 and accordingly plot the SI-regret in Figure 6, the SPR of the marginals obtained via SI-MWU in Figure 7 and the SPR of the robust averaging w_i^t in Figure 8.

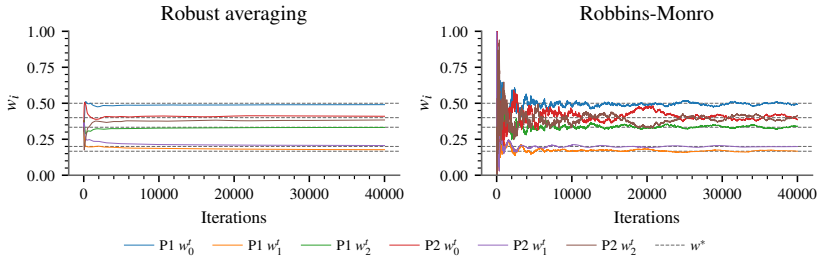


Figure 4: Learnt weights using (L) robust averaging ($\eta_t \propto 1/\sqrt{t}$ and averaging as per Equation (37)), and (R) Algorithm 2 as written ($\eta_t \propto 1/t$, no averaging), on SI-MWU output.

D.4 Effect of Stochastic Action Sets on Compute Times

In Figure 5 we show the impact of action availability on the number of iterates to solve a GSAS. We compare three different games:

1. P1 has 25 independently randomly available actions, P2 has 200 always available actions.
2. P1 has 200 independently randomly available actions, P2 has 25 always available actions.
3. Both players always have all actions available.

We use the same payoff matrix, potentially transposed, for all games. We generate the payoff matrix and ρ_1 the same as in 'Random $n \times n$ GSAS' (Definition D.2). We solve the nonstochastic game with both SI-MWU and MWU, achieving similar results. This seems to indicates that the somewhat slow rate of convergence to NE (compared to practically used algorithms such as Regret Matching and variants thereof) is an artifact of MWU as a base algorithm instead of the SI-MWU algorithm itself. This further motivates future work to study SI-variants of Regret Matching, for instance.

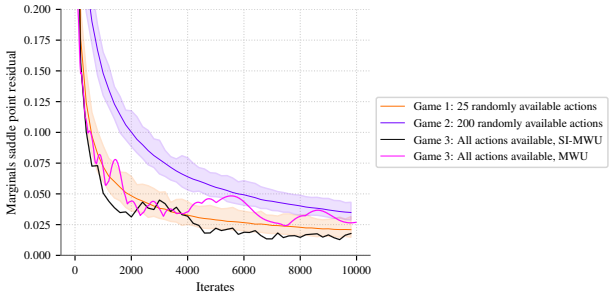


Figure 5: Saddle point residuals of the marginal distribution played by SI-MWU. We compare different 25×200 random games where either the player with 25, or 200, actions has stochastic action sets, and the nonstochastic game with the same payoff matrix solved by both SI-MWU and regular MWU, as described in D.4. We repeat the experiment for 100 different randomly generated payoff matrices (generated as in Definition D.2) and plot the average and 95% central interval.

D.5 Experiment 2: convergence of SI-MWU and Algorithm 2 for additional GSAS

In Figure 6, we show further regret plots on several additional 2p0s-GSAS: Example 3.7, biased 100×100 Rock-Paper-Scissors (Definition D.4) and biased 100×100 Matching Pennies (Definition D.5). We show the saddle point residual on those same games for the marginals in Figure 7 and of the robust averaging w_i^t in Figure 8. All results match our theoretical bounds as in Section 5.

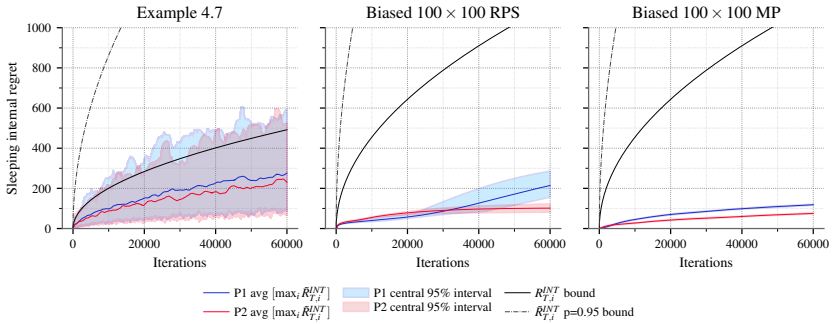


Figure 6: Sleeping internal regrets from SI-MWU for several 2p0s-GSAS. We repeated each experiment 100 times and graph for both players the average $\max_i \tilde{R}^{INT}_{T,i}$ over the repetitions, bounded by Theorem 4.4 and the central 95% interval of $\max_i \tilde{R}^{INT}_{T,I}$ over the repetitions, bounded by Proposition 4.5.

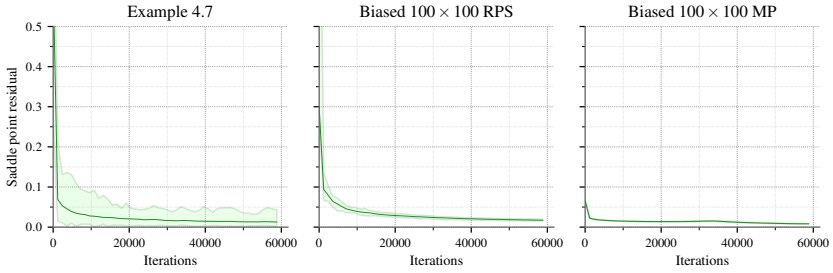


Figure 7: Saddle point residuals of the marginal distribution played by SI-MWU for several 2p0s-GSAS. We repeated each experiment 100 times and graph for both players the average and the range.

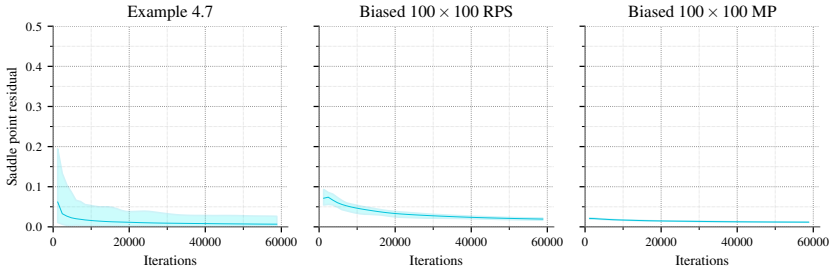


Figure 8: Saddle point residuals of the w^t_i learnt by 2 on the marginals produced by SI-MWU for several 2p0s-GSAS. We repeated each experiment 100 times and graph for both players the average and the range of the saddle point residual. For better stability of the stochastic approximation, we discard the first 500 marginals produced by SI-MWU in these plots.

E Computing Compact NE in General-Sum Games

Suppose the GSAS has ρ which is unknown, but assume that for each joint available set S^t, we can solve the sub-game $\mathcal{G}_{S^t}$ restricted to S^t. For instance, in a decentralized learning setting, each player might observe their realized action availability set S^t_i and report it to a joint controller that runs e.g. an LP solver to obtain a Nash equilibrium of the sub-game. Note that each sub-game has size at most $|A|$, and so does not suffer from the exponential dependence for solving the full GSAS. Then, Algorithm 3 provides a procedure to compute the compact Nash equilibrium vector w of the full GSAS based on the stochastic approximation approach in Algorithm 2.

Algorithm 3 Computing Compact Nash Equilibria in General-Sum Games

Initialize $\theta_i^1 \leftarrow \mathbf{1}_{|A_i|}$ for each player $i \in \mathcal{I}$;
for $t = 1, 2, \ldots, T$ **do**
  Observe the joint available set S^t;
  Compute the equilibrium $\pi^t(S^t)$ of the game $\mathcal{G}$ restricted on the available set S^t;
  for each player $i \in \mathcal{I}$ **do**
    $G_i^t(a_i) \leftarrow \pi_i^t(a_i|S^t) - \frac{\exp(\theta_i^t(a_i)\mathbb{1}\{a_i \in S_i^t\})}{\sum_{a_i' \in S_i^t} \exp(\theta_i^t(a_i'))}$, for all $a_i \in A_i$;
    $\theta_i^{t+1} \leftarrow \theta_i^t + \eta_t G_i^t$;
  end for
end for
for each player $i \in \mathcal{I}$ **do**
  $w_i = \exp(\theta_i^T)/\sum_{a_i \in A_i} \exp(\theta_i^T(a_i))$;
end for
return $w = (w_1, \ldots, w_n)$;

Proposition E.1. *Let w be the collection of weight vectors computed by Algorithm 3 applied to a general-sum GSAS $\mathcal{G}$, and let π be the corresponding strategy profile computed from w. Then, π is a Nash equilibrium of $\mathcal{G}$.*

Proof. Let $\mu_i = \frac{1}{T}\sum_{t=1}^T \pi_i^t(S^t)$ be the average strategy of player i. Since $\pi^t(S^t)$ is a Nash equilibrium of the game restricted in S^t, it follows that μ_i gives the maximum expected payoff to player i. Moreover, we have for each player i and $a_i \in A_i$ that,

$$\mu_i(a_i) = \frac{1}{T}\sum_{t=1}^T \pi_i^t(a_i|S^t) = \sum_{S \in \mathcal{S}} \rho(S)\pi_i^*(a_i|S) = \sum_{S_i \in \mathcal{S}_i} \rho_i(S_i)\bar{\pi}_i(a_i|S_i)$$

where $\pi^*(S)$ is the Nash equilibrium of the game restricted in S and

$$\bar{\pi}_i(a_i|S_i) = \sum_{S_{-i} \in \mathcal{S}_{-i}} \rho_{-i}(S_{-i})\pi_i^*(a_i|S_i, S_{-i}).$$

This implies that μ_i is implementable by $\bar{\pi}$, and therefore, by Proposition 3.2, any strategy profile that implements $\mu = (\mu_1, \ldots, \mu_n)$ is a Nash equilibrium. Moreover, it follows from Proposition 4.7 that the strategy profile π computed from w implements μ. Therefore, π is a Nash equilibrium, as desired. □

We note that as in Proposition 4.7, the convergence of Algorithm 3 to the Nash of $\mathcal{G}$ is asymptotic in nature. Similar finite-time approximations of the convergence rate can be obtained using robust stochastic approximation, as outlined in Appendix C.4.

F Optimism and Sleeping Internal Regret

A natural extension of online learning algorithms in games is to introduce 'optimism' to exploit the predictability of the game payoffs. These variants typically enjoy better regret bounds, and even last-iterate convergence to equilibria in normal-form games (Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015; Daskalakis et al., 2021; Anagnostides et al., 2022b). We investigate the behavior of using Optimistic MWU (OMWU) in place of MWU in the lower level of Algorithm 1.

Remark F.1. *The notion of optimism in stochastic bandit settings often arises in UCB-type algorithms, which were introduced initially by Lai & Robbins (1985), while Auer et al. (2002) gave a finite-time analysis of the approach. While the terminology is similar, the notion of optimism used in the bandit setting gives finer control over the exploration by constructing confidence intervals based on past samples. Applying this approach to the GSAS setting could provide improved regret bounds, though we leave this investigation to future work.*

The optimistic counterpart to the MWU step (Line 8 in Algorithm 1) is given by:

$$\tilde{q}^{t+1}(a_i \to a_i') \propto \tilde{q}^t(a_i \to a_i') \exp(-2\eta \ell^t(a_i \to a_i') + \eta \ell^{t-1}(a_i \to a_i')) \tag{46}$$

where $\ell^t(a_i \to a_i')$ is as defined in Equation (5). We will henceforth refer to the modified algorithm as SI-OMWU. In static two-player zero-sum games, one can show both polylog regret and last-iterate convergence to Nash equilibria using OMWU. In stark contrast, in our setting we obtain a lower bound on the sleeping internal regret accrued by the SI-OMWU algorithm.

Proposition F.2. *There exists a 2p0s-GSAS where SI-OMWU obtains internal sleeping regret* $R_{T,i}^{\mathsf{INT}}(a_i \to a_i') \geq \Omega(\sqrt{T})$ *for all* $a_i, a_i' \in A_i, a_i \neq a_i'$.

Proof. Consider a Matching Pennies game with action set $A_1 = A_2 = \{1, 2\}$ and payoff matrix given by

1	-1
-1	1

For simplicity, assume that Player 2 always has access to both actions. Player 1 observes action subsets $\{1\}$, $\{2\}$, and $\{1, 2\}$ with equal probability. The players are initialized randomly at a point which is not the unique, mixed Nash equilibrium. Our goal is to show that in this simple game, the SI-regret obtained by SI-OMWU grows as $\Omega(\sqrt{T})$.

In our setting, the standard RVU bound for external regret does not apply. To see why this is true, recall that the OMWU regret bound (written for Player 1 and suppressing the player's index) is given by:

$$R_T \leq \frac{D_{KL}(x^* \| x^1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\ell^t - m^t\|_*^2 - \frac{1}{8\eta} \sum_{t=2}^{T} \|x^t - x^{t-1}\|^2,$$

where m^t is the predictor of the next loss ℓ^t. In OMWU, $m^t = \ell^{t-1}$. In a zero-sum game, the term $\sum_{t=1}^{T} \|\ell^t - m^t\|^2$ is easily controlled (Syrgkanis et al., 2015), which is not the case in our setting. In particular, let $\chi^t = \mathbb{1}[S^t = \{1, 2\}]$ be a Bernoulli random variable describing if Player 1 has access to both actions or not. Since $P(\chi^t = 1, \chi^{t-1} = 0) = P(\chi^t = 0, \chi^{t-1} = 1) = 2/9$,

$$\mathbb{E}[\|\ell^t - \ell^{t-1}\|^2] = \frac{2}{9}\|\ell^t\|^2 + \frac{2}{9}\|\ell^{t-1}\|^2 \geq \frac{4}{9} = \Omega(1), \tag{47}$$

where we have used the fact that $\|\ell^t\| \geq 1$ since ℓ^t is the loss vector for Player 1 given Player 2's action realization a_2^t. Taking the sum we get $\mathbb{E}\left[\sum_{t=1}^{T} \|\ell^t - m^t\|^2\right] = \Omega(T)$, and substituting into the regret bound we get:

$$R_T \leq \frac{1}{\eta} + \eta T$$

Setting $\eta = \frac{1}{\sqrt{T}}$ yields $R_T \leq O(\sqrt{T})$. Intuitively, in a GSAS, OMWU does not enjoy improved regret upper bounds due to the increased variation in the quality of the optimistic prediction at each timestep. As per Remark F.1, more sophisticated notions of optimism from the bandit literature might be required to obtain better bounds.

Now we proceed with the construction of the lower bound on the *sleeping internal regret* of Player 1. Let N be a random variable denoting the number of times the action subset $\{1, 2\}$ appears in T rounds of play. Each χ^t is a Bernoulli trial, so N is binomially distributed with mean $\frac{T}{3}$ and total variance $\frac{2T}{9}$.

Let $T_{12} \subseteq T$ denote the rounds where actions $\{1, 2\}$ were selected. Within this subset of rounds, the sleeping internal regret compares the maximum cumulative utility for each action replacement and the actual sequence of play of Player 1:

$$R_T^{\mathsf{INT}} = \mathbb{E}\left[\max\left(\sum_{t \in T_{12}} u^t(1 \to 2), \sum_{t \in T_{12}} u^t(2 \to 1)\right)\right] - \mathbb{E}\left[\sum_{t \in T_{12}} (u^t(a^t))\right] \tag{48}$$

Assume that Player 2 converges to $(0.5, 0.5)$, the Nash equilibrium of the game. Since χ^t is independent between rounds, in expectation $\mathbb{E}\left[\sum_{t \in T_{12}} (u^t(a^t))\right] \to 0$ as $T \to \infty$. Next, note that the

first term in the RHS of Equation (48) can be equivalently written as:

$$\mathbb{E}\left[\max\left(\sum_{t\in T_{12}} u^t(1\to 2), \sum_{t\in T_{12}} u^t(2\to 1)\right)\right] = \mathbb{E}\left[\max\left(\sum_{t\in T_{12}} X^t, \sum_{t\in T_{12}} -X^t\right)\right], \tag{49}$$

where X^t is a Rademacher random variable. We do a similar trick as in the proof of Thm 3.7 in Cesa-Bianchi & Lugosi (2006). Indeed,

$$\mathbb{E}\left[\max\left(\sum_{t\in T_{12}} X^t, \sum_{t\in T_{12}} -X^t\right)\right] = \mathbb{E}\left[\left|\sum_{t\in T_{12}} X^t\right|\right], \tag{50}$$

which is the expected value of the absolute value of a $[1,-1]$ random walk. Utilizing Khintchine's inequality we can lower bound the ℓ_1-norm of the sum of Rademacher variables as follows:

$$\mathbb{E}\left[\left|\sum_{t\in T_{12}} X^t\right|\right] \geq \frac{1}{\sqrt{2}}\sqrt{N} \approx \sqrt{\frac{T}{6}}. \tag{51}$$

Combining the two terms, we obtain $R_T^{\mathsf{INT}} \geq \sqrt{\frac{T}{6}} + o(T) = \Omega(\sqrt{T})$ as required.

□