A. Supplementary material for Section 2

A.1. Correlated Equilibria

Here, we briefly introduce the concept of Coarse Correlated Equilibria. The traditional setting for this solution concept is that players receive private recommendations from a third-party mediator and must decide whether there is an incentive to deviate from these recommendations. In the context of multiagent learning, there is no mediator. However, players do have access to a shared experience over the course of learning. We will see that history can implicitly become a correlating device, and that the outcome of certain learning algorithms results in these correlated solution concepts.

Let π be a probability distribution over joint actions, i.e., $\pi \in \Delta[\mathcal{A}]$. As is done with utility functions, we will write π evaluated at $a = (a_1, \ldots, a_n)$ as both $\pi(a)$ and $\pi(a_i, a_{-i})$.

DEFINITION A.1 (CORRELATED EQUILIBRIUM (CE)). A joint probability distribution $\pi \in \Delta[\mathcal{A}]$ is a *correlated equilibrium* if, for every player i and every pair of actions $a_i, a_i' \in \mathcal{A}_i$,

$$\sum_{a_{-i} \in \mathcal{A}_{-i}} \pi(a_i, a_{-i})\, u_i(a_i, a_{-i}) \geq \sum_{a_{-i} \in \mathcal{A}_{-i}} \pi(a_i, a_{-i})\, u_i(a_i', a_{-i}).$$

The intuition is as follows. Suppose a mediator is randomly generating joint actions according to the probability distribution π, and privately recommends the generated a_i to player i. A CE is a probability distribution such that no player benefits from a unilateral deviation *conditional* on the recommendation received (i.e., play the alternative a_i' whenever the recommendation is a_i). With this definition of CE, the resulting relation between equilibrium concepts becomes: pure NE $\subseteq$ mixed NE $\subseteq$ CE $\subseteq$ CCE.

A.2. A Potential Function Characterization of Weakly Acyclic Games

Weakly acyclic games can be defined in terms of a potential-like function.

PROPOSITION A.1 (**Marden et al. [65]**). *A game is weakly acyclic if and only if there exists a (potential-like) function $\Phi : \mathcal{A} \to \mathbb{R}$ such that for every $a \in \mathcal{A}$ that is* not *a NE, there exists a player $i \in \mathcal{N}$ and action $a_i' \in \mathcal{A}_i$ such that (i) $u_i(a_i', a_{-i}) > u_i(a_i, a_{-i})$ and (ii) $\Phi(a_i', a_{-i}) > \Phi(a_i, a_{-i})$.*

The idea here is that for any action that is not a NE, there exists a player that can make a unilateral change to simultaneously improve its utility and increase the potential-like function.

B. Supplementary material for Section 3

B.1. Internal (Swap) Regret Matching

The regret underlying the regret matching algorithm of Section 3.2.5 is the *external regret*: comparisons against a single fixed alternative action a_i' across all past rounds. Hart and Mas-Colell [43] introduce a finer notion, the *swap regret*, that compares against conditional deviations.

In this learning rule, the *fitness forecast* of Player i is the *cumulative swap regret* from the current action $a_i(t)$ to other actions a_i'. More precisely, for actions a_i, a_i' let us define the swap regret of a_i with a_i' by

$$\mathrm{sr}_i^{a_i,a_i'}(t) = \frac{1}{t+1} \sum_{\tau \le t:\, a_i(\tau)=a_i} \Big[u_i\big(a_i', a_{-i}(\tau)\big) - u_i\big(a_i(\tau), a_{-i}(\tau)\big) \Big].$$

This is the average per iteration gain had the player replaced every past occurrence of action a_i with action a_i'. Note that the swap regret is an $\mathcal{A}_i \times \mathcal{A}_i$ matrix with zero diagonal entries. The *fitness forecast* for the swap regret algorithm is simply the *instantaneous positive swap regret* the fitness is $f_i^{a_i'}(t) := [\mathrm{sr}_i^{a_i(t),a_i'}(t)]^+$.

For the response mechanism, a revising player i applies a *proportional shift*: it adjusts its probability on action a_i' upward whenever there is positive swap regret from the currently played action $a_i(t)$ to a_i',

$$x_i^{a_i'}(t+1) = \begin{cases} \frac{1}{\mu}[\mathrm{sr}_i^{a_i(t),a_i'}(t)]^+ & \text{if } a_i' \neq a_i(t) \\ 1 - \sum_{a_i' \neq a_i(t)} x_i^{a_i'}(t+1) & \text{if } a_i' = a_i, \end{cases} \tag{B.1}$$

where $\mu > 0$ is a normalization constant ensuring $x_i(t+1) \in \Delta(\mathcal{A}_i)$.

THEOREM B.1 **(Swap Regret Matching and CE (Hart and Mas-Colell [43]))**. *When all players run Swap (internal) Regret Matching, the time-averaged empirical joint distribution $z(t)$ of (3.7) converges almost surely to the set of Correlated Equilibria of the underlying game.*

Response-Fitness Reformulation: For the *Proportional Shift-Instantaneous Positive Swap Regret* learning algorithm, the empirical frequency $z(t)$ converges to the set of CE for finite games.

The contrast with external RM (Theorem 3.5) is instructive. External regret minimization guarantees convergence only to the coarser CCE set, because it controls only unconditional deviations. Internal regret minimization controls *conditional* deviations, replacing action a_i with a_i' whenever a_i was played, which is precisely the incentive condition defining a Correlated Equilibrium. The hierarchy CE $\subseteq$ CCE is thus matched by a hierarchy of regret use in RM and SRM.

B.2. Exp3 Algorithm

Exp3 (Exponential-weight algorithm for Exploration and Exploitation) Auer et al. [7] adapts the Hedge response mechanism to via establishing a proximal hindsight fitness function. To construct

such an estimate, player i cannot observe $u_i(a_i', a_{-i}(t))$ for the unchosen actions. Instead, it constructs an *estimate* of the counterfactual payoff

$$\hat{u}_i^{a_i'}(t) = \begin{cases} \frac{u_i(a(t))}{x_i^{a_i(t)}(t)} & \text{if } a_i' = a_i(t), \\ 0 & \text{if } a_i' \neq a_i(t). \end{cases} \tag{B.2}$$

This estimator satisfies $\mathbf{E}[\hat{u}_i^{a_i'}(t) \mid a_{-i}(t)] = u_i(a_i', a_{-i}(t))$, so it is an unbiased replacement for the semi-bandit payoff signal. This unbiased estimator of $u_i(a_i', a_{-i}(t))$ then serves as the counterfactual payoffs. It is then only natural to use it for, e.g., a proxy for *Hindsight Fitness*. So, the fitness forecast for this algorithm is the *Proximal Hindsight Fitness* $\hat{h}_i(t) = \sum_{\tau \leq t} \hat{u}_i(\tau)$.

Given this fitness forecast, Exp3 applies a modified *logit* response $\hat{\text{lg}}_i : \mathbb{R}^{|\mathcal{A}_i|} \to \Delta(\mathcal{A}_i)$ given by

$$\hat{\text{lg}}_i(f) = (1-\gamma)\frac{\exp\left(\eta^{-1} f^{a_i}\right)}{\sum_{a_i' \in \mathcal{A}_i} \exp\left(\eta^{-1} f^{a_i'}\right)} + \frac{\gamma}{|\mathcal{A}_i|},$$

where $\eta > 0$ is the usual smoothing coefficient (temperature parameter) and $\gamma \in (0,1)$ is an exploration parameter. The uniform component $\gamma/|\mathcal{A}_i|$ ensures every action is played with strictly positive probability, preventing the denominator in (B.2) from approaching zero and causing estimator variance to explode.

THEOREM B.2 **(Auer et al. [7]).** *When all players independently employ Exp3 with appropriately tuned η and γ, the expected external regret for each player is bounded by $O(\sqrt{T|\mathcal{A}_i|\ln|\mathcal{A}_i|})$. Consequently, the time-averaged empirical joint distribution $z(t)$ converges almost surely to the set of Coarse Correlated Equilibria.*

Response-Fitness Reformulation: For the *Modified Logit response-Proximal Hindsight Fitness* learning algorithm, the empirical frequencies $z(t)$ converges to CCE with the regret $O(\sqrt{T|\mathcal{A}_i|\ln|\mathcal{A}_i|})$.

The slower convergence rate relative to semi-bandit algorithms, such as Hedge or FTRL, reflects the additional uncertainty involved with hindsight fitness estimate, not a deficiency in the response mechanism.

C. Supplementary material for Section 4

C.1. System-Level Analysis

As mentioned in Section 4.2 and illustrated in Figure 6, learning dynamics can be analyzed in isolation as open systems. The system-level properties of learning dynamics concern the *dynamic*

relationship between the evolving strategy $x_i(\cdot)$ in response to the a payoff vector trajectory $p_i(\cdot)$. This perspective adopts a two-step approach: (i) characterize the system-level implications of key properties without committing to specific learning dynamics, and (ii) verify that these properties hold for the learning dynamics of interest.

We begin with a continuous-time version of cumulative regret from (3.10). For the learning dynamics (4.1), define the cumulative regret as

$$R_i^k(t) = \int_0^t p_i^k(\tau) - x_i^\top(\tau) p_i(\tau)\, dt.$$

As in the discrete-time case, the $R_i^k(t)$ indicates the payoff that would have been garnered by a fixed strategy $\mathbf{e}^k$ as compared to the actual cumulative average payoff (assuming the same payoff trajectory $p_i(\tau)$). Learning dynamics are said to have *no-regret* if $\max_{k\in\mathcal{A}_i} R_i^k(t) \le o(t)$.

The notion of no-regret illustrates a system-level property that can be assessed for specific learning dynamics. For example, replicator dynamics have finite cumulative regret, which is even stronger than the no-regret property.

THEOREM C.1 **(Cheung and Piliouras [29]).** *For any $x_i(0)$, there exists a $C > 0$ such that*

$$\limsup_{t\to\infty} \max_{k\in\mathcal{A}_i} R_i^k(t) < C.$$

A system-level property closely associated with stability is δ-passivity. This concept was introduced in Fox and Shamma [36] and subsequently developed and extended in several publications (e.g., Abdelraouf et al. [1], Certório et al. [25], Martinez-Piazuelo et al. [69], Park et al. [79], Pavel [80] and references therein).

Let $\mathbb{P}$ denote the set of locally Lipschitz functions $p : [0,\infty) \to \mathbb{R}^{|\mathcal{A}_i|}$ with $\sup_{t\in[0,\infty)} \|p(t)\| < \infty$ and $\sup_{t\in[0,\infty} \|\dot{p}(t)\| < \infty$.

DEFINITION C.1 (δ-PASSIVITY). A learning rule is δ-passive if, for any $x_i(0)$ and $p(\cdot) \in \mathbb{P}$, there exists a $C > 0$ and $\eta \ge 0$ such that for all $t \in [0,\infty)$,

$$\int_0^t \dot{x}_i^\top(\tau)\dot{p}_i(\tau)\, d\tau \ge -C + \eta \int_0^t \dot{x}_i^\top(\tau)\dot{x}_i(\tau)\, d\tau.$$

It is strictly δ-passive if $\eta > 0$.

The system-level property of δ-passivity can be seen as complementary to contractive games (originally called stable games in Hofbauer and Sandholm [49]). Define $\mathcal{Z}_i = \left\{z \in \mathbb{R}^{\mathcal{A}_i} : \mathbf{1}^\top z = 0\right\}$ and let $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$.

DEFINITION C.2 (C^1 CONTRACTIVE GAME). A population game is C^1 contractive if for all $i \in \mathcal{N}$, the $P_i : \mathcal{X} \to \mathbb{R}^{|\mathcal{A}_i|}$ are continuously differentiable and satisfy for some $\beta \geq 0$

$$\sum_{i \in \mathcal{N}} z_i^\top \nabla P_i(x) z \leq -\beta z^\top z$$

for all $x \in \mathcal{X}$ and $z = (z_i, z_{-i}) \in \mathcal{Z}$. It is strictly C^1 contractive if $\beta > 0$.

While this definition is for a static relationship $p_i = P_i(x)$, there is a dynamic implication. Since $\dot{p}_i(t) = \nabla P_i(x(t))\dot{x}(t)$, contractive games satisfy (since the simplex is invariant, $\dot{x}_i(t) \in \mathcal{Z}_i$)

$$\sum_{i \in \mathcal{N}} \int_0^t \dot{x}_i^\top(\tau)\dot{p}_i(\tau)\, d\tau \leq -\beta \sum_{i \in \mathcal{N}} \int_0^t \dot{x}_i^\top(\tau)\dot{x}_i(\tau)\, d\tau.$$

THEOREM C.2. *For learning dynamics that are δ-passive coupled with a C^1 contractive game, if either is strict, then the trajectory time-derivatives are square-integrable, i.e., $\int_0^\infty \|\dot{x}(\tau)\|^2 d\tau < \infty$.*

The conclusion is a direct consequence of the definitions. Under δ-passive learning dynamics for a C^1 contractive game,

$$\sum_{i \in \mathcal{N}} \left(-C + \eta \int_0^t \dot{x}_i^\top(\tau)\dot{x}(\tau)\, d\tau \right) \leq \sum_{i \in \mathcal{N}} \int_0^t \dot{x}^\top(\tau)\dot{p}_i^\top(\tau)\, d\tau \leq -\beta \sum_{i \in \mathcal{N}} \int_0^t \dot{x}_i^\top(\tau)\dot{x}_i(\tau)\, d\tau$$

and so

$$\sum_{i \in \mathcal{N}} \eta \int_0^t \dot{x}_i^\top(\tau)\dot{x}(\tau)\, d\tau + \beta \sum_{i \in \mathcal{N}} \int_0^t \dot{x}_i^\top(\tau)\dot{x}_i(\tau)\, d\tau \leq |\mathcal{N}|\, C.$$

Either strict δ-passivity ($\eta > 0$) or strict contractive condition ($\beta > 0$) leads to the desired result.

This analysis again illustrates the implications of a system-level property, in this case δ-passivity, without committing to specific learning dynamics. Validating specific learning dynamics are δ-passive is separate. This can be accomplished through the discovery of a positive function (called a storage function) $S : \mathcal{X}_i \times \mathbb{R}^{|\mathcal{A}_i|} \to [0, \infty)$ such that

$$S(x_i(t), p_i(t)) - S(x_i(0), p_i(0)) \leq \int_0^t \dot{x}_i^\top(\tau)\dot{p}_i^\top(\tau)\, d\tau - \eta \int_0^t \dot{x}_i^\top(\tau)\dot{x}_i^\top(\tau)\, d\tau.$$

The construction of a suitable storage function to certify δ-passivity parallels the search for a suitable Lyapunov function to certify stability.

THEOREM C.3. *Logit dynamics, target gradient dynamics, and Smith dynamics are δ-passive. Replicator dynamics are not δ-passive.*

These results can be found in Abdelraouf et al. [1], Fox and Shamma [36], Park et al. [79].

C.2. Higher-Order Learning Dynamics

Higher-order learning modifies standard-order learning of the form $\dot{x}_i = \mathcal{F}(x_i, p_i)$ through the introduction of auxiliary states, as in

$$\dot{x}_i = \mathcal{F}(x_i, \psi), \quad \dot{y}_i = \mathcal{G}(y_i, p_i), \quad \psi = \Psi(y_i, p_i),$$

Laraki and Mertikopoulos [57] introduced a higher-order variant of replicator dynamics:

$$\dot{x}_i = \mathcal{F}(x_i, \psi_i) = \mathbf{diag}[\psi_i - x_i^\top \psi_i] x_i, \quad \dot{y}_i = p_i, \quad \psi_i = y_i.$$

These resemble standard replicator dynamics, with the payoff vector replaced with its cumulative integral. With this modification, higher-order replicator dynamics asymptotically eliminate weakly dominated strategies, which was not the case for standard replicator dynamics.

Arslan and Shamma [5] introduced anticipatory replicator dynamics, where

$$\dot{x}_i = \mathcal{F}(x_i, \psi_i) = \mathbf{diag}[\psi_i - x_i^\top \psi_i] x_i, \quad \dot{y}_i = \lambda(p_i - y_i), \quad \psi_i = p_i + \gamma\lambda(p_i - y_i)$$

and showed that this modification can enable convergence to interior NE when not possible with standard replicator dynamics (contrast with Theorem 4.5).

A major consequence of higher-order learning addresses a landmark impossibility result due to Hart and Mas-Colell [44]. For both discrete-time iterative learning and evolutionary continuous-time learning, there were universal learning dynamics that lead to no-regret. In contrast, the results on convergence to NE thus far depend on the specific coupling between learning dynamics and the game. This paper shows that non-universality fundamental to standard-order learning for the *mismatching pennies* game polymatrix game. There are three players with two-moves each, and

$$M_{12} = \begin{pmatrix} 0 & a_1 \\ 1 & 0 \end{pmatrix}, M_{23} = \begin{pmatrix} 0 & a_2 \\ 1 & 0 \end{pmatrix}, M_{31} = \begin{pmatrix} 0 & a_3 \\ 1 & 0 \end{pmatrix}$$

where $(a_1, a_2, a_3) \approx (1, 1, 1)$ and all other $M_{ij} = 0$. This game has a unique interior NE $x_1^* = \left(\frac{a_3}{1+a_3} \ \frac{1}{1+a_3}\right)^T, x_2^* = \left(\frac{a_1}{1+a_1} \ \frac{1}{1+a_1}\right)^T, x_3^* = \left(\frac{a_2}{1+a_2} \ \frac{1}{1+a_2}\right)^T$. Interestingly, player i's strategy at NE depends on the parameters of another player who is payoff irrelevant, i.e., a player whose action does not impact p_i.

THEOREM C.4 **(Hart and Mas-Colell [44]).** *There are no learning dynamics of the form (4.1) for which the NE is locally asymptotically stable for all mismatching pennies games in an open set around* $(a_1, a_2, a_3) = (1, 1, 1)$.

Hart and Mas-Colell [44] actually allowed a broader class of learning dynamics than (4.1). The essential features were that the learning dynamics are (i) standard order and (ii) "uncoupled" in the sense that the learning dynamics of one player cannot depend explicitly on the utility function of other players. Because the learning dynamics considered herein depend on a players own payoff vector, they are automatically uncoupled. This dependence is further illustrated in Figure 6, where the forward dynamics are diagonal—player i only measures its own p_i.

Higher-order dynamics, while still uncoupled, are able to circumvent this impossibility result.

THEOREM C.5 **(Shamma and Arslan [95]).** *There exist higher-order (uncoupled) target gradient dynamics and higher-order logit dynamics for which the NE is locally asymptotically stable for all mismatching pennies games in an open set around* $(a_1, a_2, a_3) = (1, 1, 1)$.

Accordingly, the barrier to convergence relies on *both* the learning dynamics being uncoupled and of standard order.

More recent work shows that convergence to interior mixed-strategy NE is generally possible using higher-order dynamics.

THEOREM C.6 **(Toonsi and Shamma [107]).** *For generic games with an isolated interior mixed-strategy NE, there exist higher-order target gradient dynamics for which the NE is locally asymptotically stable for the original game and an open set of nearby games.*

The implication for mixed-strategy NE is significant, as there are various impossibility results addressing convergence of learning dynamics to such equilibria (Gaunersdorfer and Hofbauer [40], Jordan [52], Vasin [110]).