Evolutionary Dynamics and
Φ-Regret Minimization in Games

Georgios Piliouras   GEORGIOS@SUTD.EDU.SG
Singapore University of Technology and Design
Singapore

Mark Rowland   MARKROWLAND@DEEPMIND.COM
Shayegan Omidshafiei   SOMIDSHAFIEI@DEEPMIND.COM
Romuald Elie   RELIE@DEEPMIND.COM
Daniel Hennes   HENNES@DEEPMIND.COM
Jerome Connor   JEROMECONNOR@DEEPMIND.COM
Karl Tuyls   KARLTUYLS@DEEPMIND.COM
DeepMind
London, United Kingdom

Abstract

Regret has been established as a foundational concept in online learning, and likewise has important applications in the analysis of learning dynamics in games. Regret quantifies the difference between a learner's performance against a baseline in hindsight. It is well-known that regret-minimizing algorithms converge to certain classes of equilibria in games; however, traditional forms of regret used in game theory predominantly consider baselines that permit deviations to deterministic actions or strategies. In this paper, we revisit our understanding of regret from the perspective of deviations over partitions of the full *mixed* strategy space (i.e., probability distributions over pure strategies), under the lens of the previously-established Φ-regret framework, which provides a continuum of stronger regret measures. Importantly, Φ-regret enables learning agents to consider deviations from and to mixed strategies, generalizing several existing notions of regret such as external, internal, and swap regret, and thus broadening the insights gained from regret-based analysis of learning algorithms. We prove here that the well-studied evolutionary learning algorithm of replicator dynamics (RD) seamlessly minimizes the strongest possible form of Φ-regret in generic 2×2 games, without any modification of the underlying algorithm itself. We subsequently conduct experiments validating our theoretical results in a suite of 144 2×2 games wherein RD exhibits a diverse set of behaviors. We conclude by providing empirical evidence of Φ-regret minimization by RD in some larger games, hinting at further opportunity for Φ-regret based study of such algorithms from both a theoretical and empirical perspective.

1. Introduction

Understanding the behavior of learning dynamics in games is a fundamental problem studied in game theory, online learning theory, dynamical systems, and multiagent systems. Numerous works have been developed in the area that describe connections between these distinct fields (Bloembergen et al., 2015; Celli et al., 2020; Cesa-Bianchi and Lugosi, 2006; Chang, 2007; Galstyan, 2013; Gatti and Restelli, 2016; Klos et al., 2010; Nisan et al., 2007; Srini-

vasan et al., 2018; Tuyls and Parsons, 2007; Tuyls et al., 2003, 2006; Vlatakis-Gkaragkounis et al., 2020; Wunder et al., 2010). Given this wealth of prior work, a natural question emerges: have we converged to a more-or-less complete mathematical language that allows us to accurately describe the behavior of learning dynamics (at least in simple small games), or are there remaining characteristics of these dynamics that are not yet well understood or fully described by existing concepts?

Arguably, one of the most important concepts in this area is that of *regret* (Auer et al., 1995; Chang and Kaelbling, 2005; Freund and Schapire, 1999; Kleinberg et al., 2009; Monnot and Piliouras, 2017; Roughgarden, 2009). Regret is a basic definition of online learning with numerous variants, each with specific associated properties. Regret minimization forms the basis of a key class of algorithms for learning in games, and is concurrently central to the more general field of online sequential prediction (Cesa-Bianchi and Lugosi, 2006; Chang, 2007). Informally, regret is defined as the difference in the cumulative performance of an agent against a baseline, which typically allows hindsight deviations (or swaps) of one or more of the agent's actions with alternative actions. There exist various notions of regret, ranging from basic forms such as external regret (Hannan, 1957), which measures a learner's performance against the best fixed action in hindsight, to stronger variants such as internal regret (Foster and Vohra, 1998) and swap-regret (Blum and Mansour, 2007a), where the deviating action of the agent is allowed to be a function of their originally chosen action. These notions of regret have typically been defined within the context of single-agent online learning, and subsequently used to analyze behavior of decision-making in multiagent games. The benefit of regret-minimizing algorithms, when applied to games, is that they typically yield time-average convergence to a corresponding set of equilibria. For example, in general-sum games, (external) regret minimization by all players yields time-average convergence to coarse correlated equilibria (Nisan et al., 2007); in two-player zero-sum games, this yields convergence to Nash equilibria, and has driven key successes towards solving games such as poker and offered other important insights in extensive-form games (Brown and Sandholm, 2018; Brown et al., 2020; Farina et al., 2021a; Moravčík et al., 2017; Morrill et al., 2021a,b; Sandholm, 2010a; Zinkevich et al., 2007).

Despite these insights, traditional notions of regret such as those described above are, at times, too general to provide strong guarantees for certain learning dynamics in games. Even in zero-sum games, for example, standard regret-minimizing dynamics such as multiplicative weights update (Arora et al., 2012b) and its well-known continuous-time limit counterpart, the replicator dynamics (RD), can be non-convergent or even chaotic in their real-time behaviors (Bailey and Piliouras, 2018; Cheung and Piliouras, 2019, 2020; Piliouras and Shamma, 2014; Sato et al., 2002), despite their time-average convergence. Emergence of chaos can even materialize in simple congestion games (Chotibut et al., 2020a; Palaiopanos et al., 2017), and perhaps surprisingly, such behaviors can occur despite these algorithms' time-average convergence to equilibria (Freund and Schapire, 1999). In fact, such prototypical learning algorithms can be non-equilibrating even in the smallest of environments, 2×2 games with two agents and two strategies, e.g., Matching Pennies (Bailey and Piliouras, 2019; Chotibut et al., 2020b; Papadimitriou and Piliouras, 2016). Overall, the behavior of well-known dynamics even in small games is diverse and non-trivial, and clearly cannot be fully inferred from standard regret analysis alone.

The clear gap between traditional regret theory and the observed empirical behaviors of these well-studied algorithms hints at the enticing possibility of using stronger notions of regret to better understand dynamical behaviors at finer levels of granularity. As it is well-known that regret-minimizing algorithms must, in general, randomize (i.e., output probability distributions over actions, rather than deterministic actions), such a stronger notion could consider a regret definition where agents may condition their deviating behavior not merely on deterministic actions (e.g., as in traditional notions of regret), but using more refined deviation functions. In investigating this, a line of prior work has analyzed a stronger concept known as Φ-regret (Gordon et al., 2008; Greenwald and Jafari, 2003; Hong, 2008; Stoltz and Lugosi, 2007), which encapsulates several more general classes of deviations in contrast to traditional regret notions. Related notions has attracted recently a lot of attention in extensive-form games as well (Farina et al., 2021a,b; Morrill et al., 2021a,b). However, as later detailed, while some of these works have introduced algorithms for Φ-regret minimization, such algorithms have either been investigated only for specific and simple classes of Φ-regret, or apply to more general classes of games albeit being significantly more intricate (e.g., require more book-keeping and are more difficult to implement).

In this paper, our primary contribution is to establish a link between Φ-regret and the simple and well-studied evolutionary dynamics algorithm of RD. RD is arguably one of the most well studied dynamics in evolutionary game theory Sandholm (2010b); Weibull (1997) and due to its well known connection to Multiplicative Weights Updates (Arora et al., 2012b; Kleinberg et al., 2009) enjoys strong external regret guarantees (Cheung and Piliouras, 2021; Kwon and Mertikopoulos, 2017; Mertikopoulos et al., 2018) making it an natural candidate for producing powerful Φ-regret guarantees as well. Our key theoretical result is that in general classes of 2×2 games, RD in self-play seamlessly minimizes the strongest possible notion of Φ-regret, without any modification of the underlying learning dynamics. Informally, this result implies that an RD learner, in self-play, attains on time-average at least the value of the game, even if time-averaging is applied to specific recurrent parts of the trajectory as defined by the agent's mixed strategy. Moreover, we prove that there exist algorithms that minimize external regret, which exhibit unbounded strong swap regret in some 2×2 games. We further ground our theoretical results in empirical analysis focusing on two-player games, introducing a version of Φ-regret we denote 'mosaic regret', which is more amenable to empirical implementation. We illustrate mosaic regret-convergence under RD in a broad range of 144 2×2 games (Bruns, 2015), with widely varying characteristics (e.g., fully cooperative games, social dilemmas, cyclical games, etc.). Finally, we conclude by showing empirical evidence that RD minimizes mosaic regret in some larger games, hinting at future avenues of further exploration.

The remainder of the paper is structured as follows. In Section 2, we overview the necessary preliminaries for establishing our theoretical results. In Section 3, we delve into the Φ-regret framework and related concepts targeted in the paper. Following this, we establish our theoretical results in Sections 4 and 5, and subsequently validate them empirically in Section 6. Finally, we conclude with key discussion points and takeaways in Section 7.

2. Preliminaries

We first review preliminaries related to game theory and online learning algorithms.

2.1 Game Theory

We study two-player normal-form games, where the first (resp., second) player has access to a finite set of *pure* strategies $\mathcal{A}^1$ (resp., $\mathcal{A}^2$). In two-player games, players 1 and 2 are also referred to as the row and column players, respectively. The joint *mixed* strategies of the players are denoted by (x, y), where $x \in \Delta(\mathcal{A}^1)$ and $y \in \Delta(\mathcal{A}^2)$. We will denote the probability that the first (resp., second player) assigns to their i-th strategy (resp., j-th strategy) as x_i (resp., y_j). The player payoffs are, respectively, specified by matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times m}$, where n and m are the number of strategies available to each player. Let a_{ij}, b_{ij} represent the payoff entries of the respective matrices. The respective utilities received by the players are $x^\top A y$ and $x^\top B y$. In a zero-sum (resp., coordination) game, players receive payoffs $A = -B$ (resp., $A = B$). Given strategy profile (x, y), the best response for each player is the strategy that maximizes their utility against the other player's current strategy. The *best response dynamics* arise when players iteratively update their policy to their best response.

In game theory, the Nash equilibrium has been well-established as a solution concept of interest, and is defined as follows.

Definition 2.1. *(Nash, 1950) A mixed strategy profile $(x_*, y_*) \in \Delta(\mathcal{A}^1) \times \Delta(\mathcal{A}^2)$ is a Nash equilibrium (NE) if*

$$x_*^T A y_* \geq x^\top A y_* \quad \forall x \in \Delta(\mathcal{A}^1) \quad and \quad x_*^T B y_* \geq x_*^T B y \quad \forall y \in \Delta(\mathcal{A}^2). \tag{1}$$

In other words, the players are simultaneously in best response with one another when their profiles constitute an NE.

Several weaker notions of equilibria are also important in game theory, and bear a close relationship to regret-minimizing strategies. In preparation for those definitions, it will be useful to introduce some notation that allows for us to work with arbitrary correlated probability distributions over $\mathcal{A} = \mathcal{A}^1 \times \mathcal{A}^2$. Let $z \in \Delta(\mathcal{A})$ be such a probability distribution. Let z_{ij} be the probability assigned to the outcome where the first (resp., second) player choose strategy i (resp., j). Let $z(1|\cdot)$ (resp., $z(2|\cdot)$) be the marginal probability of the first (resp., second) player. We denote by $z(2|i) \in \Delta(\mathcal{A}^2)$ the conditional distribution of the second player's strategy given the first player's strategy is i, and similarly define $z(1|j)$. Finally, we denote by $e_i \in \Delta(\mathcal{A}^1)$ the distribution putting mass 1 on $i \in \mathcal{A}^1$, and similarly define e_j. Given these definitions, we next introduce several additional notions of equilibria.

Definition 2.2. *(Aumann, 1974) A distribution over joint strategies $z \in \Delta(\mathcal{A}^1 \times \mathcal{A}^2)$ is a correlated equilibrium (CE) if*

$$(e_i)^\top A z(2|i) \geq x^\top A z(2|i) \quad \forall i \in \mathcal{A}^1, x \in \Delta(\mathcal{A}^1) \ and \tag{2}$$

$$z(1|j)^\top B e_j \geq z(1|j)^\top B y \quad \forall j \in \mathcal{A}^2, y \in \Delta(\mathcal{A}^2). \tag{3}$$

Definition 2.3. *(Moulin and Vial, 1978) A distribution over joint strategies $z \in \Delta(\mathcal{A}^1 \times \mathcal{A}^2)$ is a coarse correlated equilibrium (CCE) if*

$$\sum_{i,j} a_{ij} z_{ij} \geq x^\top A z(2|\cdot) \quad \forall x \in \Delta(\mathcal{A}^1) \ and \tag{4}$$

$$\sum_{i,j} b_{ij} z_{ij} \geq z(1|\cdot)^\top B y \quad \forall y \in \Delta(\mathcal{A}^2). \tag{5}$$

It is important to note that the set of CCE is a superset of the CE, which itself is a superset of NE (i.e., NE ⊆ CE ⊆ CCE). An intuitive way to establish this set of inclusions is to define each of these solution concepts via allowable sets of joint distributions and classes of allowable deviating policies under which no player can strictly improve their payoff. From this perspective, a NE is a *product* of (mixed) strategies such that no player can deviate to another (mixed) strategy and strictly increase their payoff. The set of CCE can be defined equivalently merely by removing the restriction that the joint distribution of the two players has to be a product of each of the player's marginal distributions. Hence, CCE is a superset of NE. Finally, if the players' initial joint distribution is not a mixture, one can define strictly more powerful deviating strategies by allowing a player's deviating strategy to depend on their realized strategy. This is the case for CE, where the set of allowable deviating strategies is strictly more expansive, and yet none of them can strictly improve the payoffs of the players; thus, the set of CE is more restrictive than CCE, while generalizing NE as, once again, it allows for correlated joint distributions.

2.2 Replicator Dynamics

In this paper, we seek to investigate the connection of more expansive notions of regret to simple, well-studied learning algorithms. For these purposes, we focus on the replicator dynamics (RD), a standard and well-studied model defining the evolution of strategic, interacting individuals under biologically-inspired mechanisms (Schuster and Sigmund, 1983; Taylor and Jonker, 1978). In the two-player setting of interest, the time-evolution of player strategies is described by RD as follows,

$$\dot{x}_i = x_i \left((Ay)_i - x^\top Ay \right) \quad \forall i \in \mathcal{A}^1 \text{ and} \tag{6}$$

$$\dot{y}_j = y_j \left((x^\top B)_j - x^\top By \right) \quad \forall j \in \mathcal{A}^2. \tag{7}$$

RD is the continuous-time variant of the well-known multiplicative weights update (MWU) meta-algorithm (Arora et al., 2012b; Kleinberg et al., 2009), and the seminal dynamics in the areas of mathematical evolution, biology, ecology, and evolutionary game theory (Hofbauer and Sigmund, 1998; Weibull, 1997). In recent years, RD has enjoyed a particularly strong surge in applications to learning in multiplayer games (Boone and Piliouras, 2019; Flokas et al., 2020; Hennes et al., 2020; Lanctot et al., 2019; Nagarajan et al., 2020; Omidshafiei et al., 2019; Papadimitriou and Piliouras, 2019; Sanders et al., 2018; Skoulakis et al., 2021; Sorin, 2020). Despite its algorithmic simplicity, RD is well-known to minimize external regret (a concept later detailed in Section 3.1), thus yielding time-average convergence to a coarse correlated equilibrium (Mertikopoulos et al., 2018; Sorin, 2009).

2.3 Online Sequential Prediction

The central problem of online sequential prediction in continuous time is specified by the interaction of a player with finite action set $\mathcal{A}$ and an environment at each time $t \in [0, \infty)$.[1] At time t, the player selects a distribution over actions $x^t \in \Delta(\mathcal{A})$, and a utility function

---

1. We emphasize that $\mathcal{A}$ is the action set of a single player, analogously with $\mathcal{A}^1$ and $\mathcal{A}^2$ in the game-theoretic setup above.
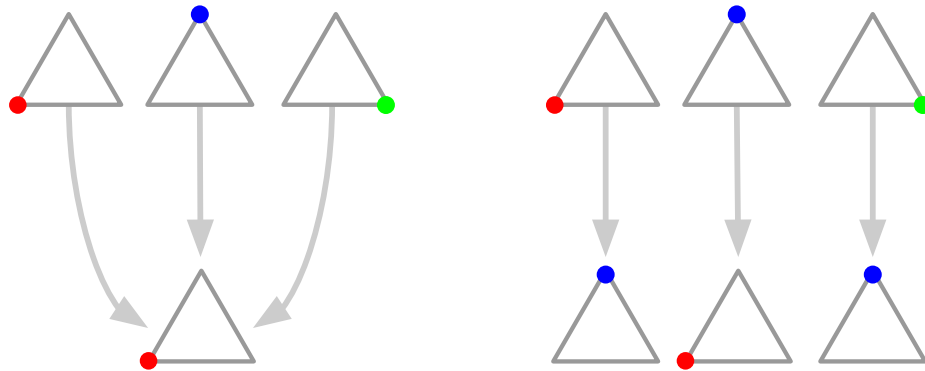
(a) Example external regret deviations.　　(b) Example swap regret deviations.

Figure 1: Strategy deviations associated with different forms of regret. This example considers a simple online learning setting with three actions (i.e., pure strategies) available to the player, illustrated via the simplices in each row. (a) External regret considers deviations of all three possible actions to a single, fixed, deterministic action in hindsight. (b) Swap regret considers a function independently swapping *each* of the player's actions with an alternative fixed action.

$u^t \in \mathbb{R}^{\mathcal{A}}$ is revealed.[2] The player's expected instantaneous utility is $\langle x^t, u^t \rangle$, and the player is shown the entire utility vector u^t, which it may use in deciding how to act in subsequent interactions. After interacting up to time $T > 0$, the cumulative utility experienced by the player is $\int_0^T \langle x^t, u^t \rangle \mathrm{d}t$. It is difficult to judge how well the player has done in selecting its actions on the basis of this cumulative utility alone; the quality of the performance depends on whether there were other actions available that would have yielded significantly higher utility. This is formalized by judging the player's performance based on its *regret*, a common means of studying the performance of such algorithms in both discrete and continuous time (Banerjee and Peng, 2004; Blum and Mansour, 2007b; Cesa-Bianchi and Lugosi, 2006; Harris, 1998; Kwon and Mertikopoulos, 2017; Mertikopoulos et al., 2018; Shoham and Leyton-Brown, 2008; Sorin, 2009; Zinkevich, 2003).

3. From Traditional Regret Minimization to Φ-Regret and Mosaic Regret

We next lay the foundations for our theoretical results, by overviewing a spectrum of no-regret algorithms: from the more traditional forms of regret to the more general Φ-regret framework, and our introduced notion of mosaic regret.

3.1 External and Swap Regret

Informally, regret quantifies whether a player could have done better by using an alternative method for picking actions throughout the interaction. One of the most common variants

---

2. The typical language of online learning uses loss vectors, $\ell^t = -u^t$. Here, we use utilities to match the focus on payoffs, rather than losses, which is common in the game theory literature.

studied is *external regret*, the expected improvement in performance that could have been achieved by sticking with a single action throughout all interactions during the time interval $[0, T]$. Figure 1(a) provides an illustrative example of the action deviations considered when computing external regret. This example considers a simple online learning setting with three actions available to the player (illustrated via each of the 2-simplices), with external regret considering deviations of all three actions to a single, fixed action in hindsight. External regret is mathematically defined by

$$\max_{a \in \mathcal{A}} \int_0^T u_a^t \mathrm{d}t - \int_0^T \langle x^t, u^t \rangle \mathrm{d}t \,, \tag{8}$$

where u_a^t is the utility of action $a \in \mathcal{A}$ at time t. If a player is able to attain $o(T)$ external regret, over bounded utility sequences ($u^t \in [0,1]^{\mathcal{A}}$, for example), then the player's strategy is said to minimize external regret, or is simply *regret-minimizing*. Intuitively, the suboptimality of the player's decision at each timestep, relative to the best constant action in hindsight, becomes vanishingly small, no matter what sequence of utilities $(u^t)_{t \in [0,T]}$ are yielded by the environment.

Regret in games. Regret minimization is central to many algorithms for computing equilibria in game theory, due to the close relationship between the notion of an alternative action, and the game-theoretic notions of strategy deviations that feature in the definitions of equilibria in Section 2.1. As such, one can consider *actions* in the sense of online learning as being synonymous with (pure) *strategies* in game theory; we henceforth use the latter terminology for simplicity. For instance, consider a two-player game as specified in Section 2.1. Let us cast the problem the players face in playing the game as an online sequential prediction problem, focusing on the first player in the following description. The first player's strategy set is $[n]$, the set of strategies in the game. The utility vector u^t for this player at time t is given by Ay^t, where y^t is the strategy selected by the second player at time t.

A well-known folk-theorem implies that if both players employ algorithms that minimize external regret to select their strategies, the players' time-average behavior (i.e., the joint strategy $T^{-1} \int_0^T (x^t, y^t) \mathrm{d}t$) is guaranteed to converge to the set of coarse correlated equilibrium (Hart and Mas-Colell, 2000; Roughgarden, 2016; Young, 2004). Further, if the game is zero-sum, then the product of the marginals of their individual time-averaged strategies converges to the set of Nash equilibria at the same rate (Freund and Schapire, 1999; Nisan et al., 2007; Young, 2004). A similar relationship holds between the set of correlated equilibria, and the stronger notions of regret described below (Hart and Mas-Colell, 2000).

Broader deviation classes. There exist more general notions of regret that compare a player's behavior against a wider class of baselines than just those that deviate to a single, fixed strategy throughout time. One such alternative, *swap regret* (Blum and Mansour, 2007a), permits deviations involving the player using strategy $b \in \mathcal{A}$ every time they had selected strategy $a \in \mathcal{A}$. Figure 1(b) illustrates swap regret in our earlier example, where now each of the three possible pure strategies may be independently deviated to a different one. The notion of swap regret is formalized through *swap* functions $F : \mathcal{A} \to \mathcal{A}$ that can

be lifted to a function $\bar{F} : \Delta(\mathcal{A}) \to \Delta(\mathcal{A})$

$$\bar{F}(x)_b = \sum_{\substack{a \in \mathcal{A} \\ F(a)=b}} x_a \,,$$

for all $x \in \Delta(\mathcal{A})$. The swap regret is then defined to be

$$\max_{F:\mathcal{A} \to \mathcal{A}} \int_0^T \langle \bar{F}(x^t), u^t \rangle \mathrm{d}t - \int_0^T \langle x^t, u^t \rangle \mathrm{d}t \,.$$

Players interacting in a two-player game using algorithms that minimize swap regret are guaranteed to converge to the set of correlated equilibria in time-average, a stronger notion that the coarse correlated equilibrium guaranteed by algorithms that minimize external regret. A slightly weaker notion than swap regret is that of *internal regret*, which restricts the swap functions $F : \mathcal{A} \to \mathcal{A}$ that are lifted to deviations on the simplex to take the form $F(a) = a$ for all but one $a \in \mathcal{A}$. However, note that in the special case of zero-sum games, internal and swap regret minimization algorithms do not offer stronger guarantees, as external regret minimization algorithms already guarantee convergence to a Nash equilibrium in the time-average.

3.2 Φ-regret Framework

In swap regret, each swap function $F : \mathcal{A} \to \mathcal{A}$ is lifted to a function $\bar{F} : \Delta(\mathcal{A}) \to \Delta(\mathcal{A})$, representing an admissible deviation in the algorithm's past strategies. Notably, each such $\bar{F}$ is *affine* in the input strategy. This is also true of the deviation functions $\bar{F} : \Delta(\mathcal{A}) \to \Delta(\mathcal{A})$ permitted under external and internal regret. A natural question to pose is whether there exist learning algorithms that minimize regret against even broader classes of deviation functions.

To better intuit the possibilities, let us consider an alternative view of the deviations permitted by various regret notions in Fig. 2. As in our earlier example, we consider a simple 3-strategy game in this instance. However, we now also visualize the full spectrum of mixed strategies, with the color of each mixed strategy in the interior of the simplex corresponding to its composition of the three underlying pure strategies (red, green, and blue). For instance, Fig. 2(a) visualizes the three possibilities of deviations permitted by external regret, each respectively mapping all mixed strategies to either the red, green, or blue pure strategy. Likewise, in the swap regret case shown in Fig. 2(b), deviations can correspond to affine transformations of the input strategies. Here, the deterministic swap function F yields a mixed swap function $\bar{F}$, mapping each input mixed strategy to a distinct output mixed strategy. Naturally, as swap regret generalizes external regret, deviations permitted in the latter are also possible under the swap regret case (e.g., as shown in example deviation #3 in Fig. 2(b)).

Φ-regret. A previous line of work (Gordon et al., 2008; Greenwald and Jafari, 2003; Stoltz and Lugosi, 2007) introduces the concept of Φ-regret, expanding the space of deviations permitted compared to traditional regret notions.[3] For our purposes, Φ-regret is defined

---

3. Φ-regret as introduced by Greenwald and Jafari (2003) applied only to linear deviation functions over $\Delta(\mathcal{A})$. Later work (Gordon et al., 2008; Stoltz and Lugosi, 2007) applied the concept much more generally
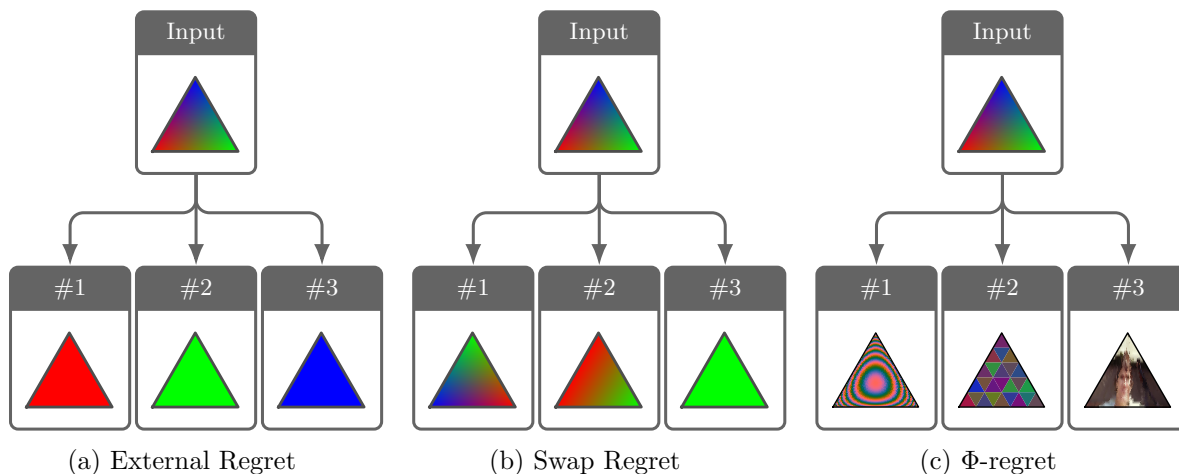
Figure 2: Mixed-strategy view of permissible deviations under various forms of regret. We visualize here a simple 3-strategy game, where the color of each mixed strategy in the simplex interior corresponds to the combination of pure strategies (red, green, and blue) that compose it. (a) Under external regret, permitted deviations map all mixed strategies to either the red, green, or blue deterministic strategy. (b) Swap regret deviations can correspond to affine transformations of the input strategies, generalizing external regret. (c) Under Φ-regret, the input space may be partitioned in a multitude of ways. For instance, in example #1, the simplex is discretized into partitions, each capturing mixed strategies with a certain range of KL-divergences from the center of the simplex. Likewise, in example #2, a discrete set of deviation partitions is defined in a granular manner enabling the output mappings visualized. This definition allows for arbitrarily complex mappings, as illustrated by example #3, which reproduces an image from the three primary strategies (colors).

by first specifying a set of deviation functions $\Phi \subseteq \{\bar{F} : \Delta(\mathcal{A}) \to \Delta(\mathcal{A})\}$, and then defining the corresponding Φ-regret as

$$\max_{\phi \in \Phi} \int_0^T \langle \phi(x^t), u^t \rangle \mathrm{d}t - \int_0^T \langle x^t, u^t \rangle \mathrm{d}t \,.$$

An algorithm is said to minimize Φ-regret if the above expression is $o(T)$; that is, it grows sub-linearly with time.

Figure 2(c) visualizes example deviations permitted under Φ-regret, illustrating that the input space may be partitioned in a multitude of ways under the above definition. For instance, in example #1 of Fig. 2(c), partitions are defined as a function of KL-divergence from the center of the simplex. Likewise, in example #2 of Fig. 2(c), deviations are defined in a manner enabling the highly granular output mappings visualized. Theoretically, this definition would allow for any arbitrarily complex mapping, as illustrated by example #3 of Fig. 2(c), which reproduces an image from the three primary strategies (colors).

Mathematically, the Φ-regret framework generalizes the concepts of external regret, internal regret, and swap regret. Additional specific instances of Φ-regret, however, can be

---

to convex games on compact action spaces. The definition we use in this paper is a specific case of this notion, applied to the mixed extension of normal-form games.

considerably stronger than these notions of regret, as any continuous function on $\Delta(\mathcal{A})$ can be uniformly approximated by piecewise affine functions.

Strong Swap Regret. The strongest version of Φ-regret takes Φ to be as large as possible; the set of all measurable functions from $\Delta(\mathcal{A})$ to itself. We refer to this notion as *strong swap regret*, as it is defined explicitly as a mapping over the mixed-strategy simplex and may, thus, be *non-affine* (in contrast to the traditional notion of swap regret, which is first defined over pure strategies before being lifted to the mixed space). Past work has led to the development of algorithms that are able to minimize this notion of regret, which is an extremely strong property, although the algorithms necessarily to do so are rather intricate and impractical to implement (Stoltz and Lugosi, 2007). One of the central findings of our work is that in certain circumstances, the straightforward algorithm of replicator dynamics actually minimizes strong swap regret.

3.3 Mosaic Regret: A Practical Strong Swap Regret Variant

In moving from theory to experiments, strong swap regret can no longer practically be minimized due to time discretization effectively preventing the learning dynamics from revisiting specific mixed strategies. It is, therefore, useful to consider a slight weakening of this notion to enable practical evaluation, which we define as follows.[4]

Definition 3.1 (Mosaic regret)**.** *A learning algorithm that produces a strategy process* $(x^t)_{t \geq 0}$ *in response to a utility process* $(u^t)_{t \geq 0}$ *is said to minimize mosaic regret if, for any finite measurable partition* $\Omega = \{\Omega_1, \ldots, \Omega_K\}$ *of* $\Delta(\mathcal{A})$, *the algorithm has vanishing regret relative to all deviation strategies that are affine on the elements of the partition* Ω. *Mathematically, given a finite measurable partition* Ω *of* $\Delta(\mathcal{A})$, *let* $D(\Omega) = \{G : \Delta(\mathcal{A}) \to \Delta(\mathcal{A}) \mid G$ *affine on each* $\Omega_k\}$ *be the corresponding set of piecewise-affine functions on* $\Delta(\mathcal{A})$. *Then the algorithm minimizes mosaic regret if*

$$\max_{G \in D(\Omega)} \int_0^T \langle G(x^t), u^t \rangle \mathrm{d}t - \int_0^T \langle x^t, u^t \rangle \mathrm{d}t = o(T)$$

over measurable, bounded utilities $u \in ([0,1]^{\mathcal{A}})^{[0,\infty)}$, *for each finite partition* Ω *of* $\Delta(\mathcal{A})$.

Mosaic regret also encompasses many standard existing notions of regret. Minimization of mosaic regret, for example, encompasses minimization of external regret through the singleton partition $\Omega = \{\Delta(\mathcal{A})\}$ and constant deviation maps. Internal and swap regret are also encompassed through the singleton partition, and various classes of affine maps. We shall return to this introduced notion of mosaic regret in our later experiments in Section 6.

4. Analysis of Replicator Dynamics under Φ-Regret

We next present theoretical results pertaining to the dynamics of games wherein players use RD. We focus our attention on two-player 2×2 normal-form games, which, despite their

---

4. For related reasons, Gordon et al. (2008) introduce a similarly-weakened notion known as finite-element Φ-regret, which restricts deviations to be continuous, and affine on each piece of a finite partition. Our introduced notion of mosaic regret is a slight strengthening of finite-element Φ-regret, allowing for non-continuous deviations.

size, form an important class capturing numerous canonical games traditionally used for closed-form analysis of regret-minimizing algorithms (Bruns, 2015; Guyer and Rapoport, 1972; Klos et al., 2010; Rapoport, 1966; Robinson and Goforth, 2005). We first begin by deriving preliminary results, subsequently establishing that RD minimizes strong swap regret in all classes of 2×2 games. Finally, we conclude by revisiting strong swap regret in the context of online learning, and in comparison to the standard notion of external regret.

We seek to establish the key result that strong swap regret is minimized in all generic 2×2 games when both players are using RD. The notion of genericity that we consider here is that we only allow for games where, for each pure strategy of their opponent, each player has a unique (pure) best response as well as that for both payoff matrices A, B it must hold that $a_{11} - a_{12} - a_{21} + a_{22} \neq 0$ and $b_{11} - b_{12} - b_{21} + b_{22} \neq 0$.[5] Clearly, these properties are satisfied by all but a zero-measure set of games from the space of all 2×2 games.

To prove the minimization of strong swap regret as outlined above, we consider the three possible classes of such games: (I) those with no pure Nash equilibrium (that have a unique interior, i.e., fully mixed, Nash equilibrium), (II) those with a unique pure Nash equilibrium, and (III) those with (at least) two pure Nash equilibria on the boundary (due to our genericity assumptions we can only have at most two pure Nash equilibria). From the perspective of a *best response graph*[6] that connects pure strategy profiles via directed edges from one to another by a single player best-responding, these cases respectively correspond to: (I) the best response graph is a cycle through the four pure strategy outcomes, (II) the best response graph has a single sink that is a pure (strict) Nash equilibrium, and (III) the best response graph has two sinks/pure (strict) Nash equilibria. In case (III), due to our genericity assumption, both players choose a different pure strategy in each pure Nash equilibrium, which implies the existence of an isolated fully mixed Nash equilibrium as well.

4.1 Equivalent and Rescaled Games

Before proving the result that RD minimizes strong swap regret in generic 2×2 games, we will show in this section that in order to prove case (I) it is sufficient to prove this result for the case of rescaled zero-sum games with a unique interior Nash equilibrium, whereas to prove case (III) it is sufficient to consider the case of rescaled coordination games with a unique interior Nash equilibrium. Let us first define these notions of equivalent and rescaled games below.

---

5. A natural sufficient condition for the uniqueness of best response strategies is that all payoff entries are distinct. It is easy to see that in 2×2 games under uniqueness of best responses, other useful generic properties follow, such as a finite number of possible Nash equilibria. Indeed, if, given a pure strategy of the opponent, each strategy has a unique best response then there cannot exist any mixed Nash equilibrium where exactly one of the two agents is randomizing. So the only possibility for having a continuum of Nash equilibria is in its interior. However, given any two interior Nash equilibria, any linear combination of them is also a Nash equilibrium, where in fact both agents are strictly indifferent between their two strategies. This line of equilibria intersects the boundary, implying the existence of a strategy profile where at least one of the agents plays a pure strategy and their opponent is indifferent between their two strategies, reaching a contradiction.

6. A best response graph is a directed graph with one vertex for each strategy profile. It has a directed edge between two vertices/profiles if and only if this transition corresponds to a best response of some agent. For more discussion see (Goemans et al., 2005; Omidshafiei et al., 2019).

Definition 4.1 (Equivalent games up to column/row shifts)**.** *Let $G = (A, B)$ be a two player game with payoff matrices A and B. We say that a game $G' = (A', B')$ is equivalent to G up to column/row shifts if there exists (c_1, c_2, d_1, d_2) such that*

$$a'_{ij} = a_{ij} + c_j, \quad b'_{ij} = b_{ij} + d_i \quad \text{for all } (i, j) \in \{1, 2\} .$$

We then write $(A, B) \sim (A', B')$ to denote equivalence up to column/row shifts.

A game is called *rescaled zero-sum* (resp., *rescaled coordination game*) if there exist a payoff matrix C and a negative (resp., positive) constant c such that $(A, B) \sim (C, cC)$. Observe that since the RD update expression for the row player can be written as

$$\dot{x}_i = x_i \sum_j x_j \left((Ay)_i - (Ay)_j \right) ,$$

the row player's RD trajectories only depend on payoff differences, which remain invariant after column payoff shifts. Hence, shifting a game to an equivalent version does not affect RD, implying the following remark.

Remark 4.2. *The orbits in RD of two equivalent games are identical.*

The following simple lemma (see also Hofbauer and Sigmund 1998) establishes that every generic 2×2 game is RD trajectory equivalent to either a rescaled zero-sum game or a rescaled coordination game.

Lemma 4.3. *Every generic 2×2 game is either a rescaled zero-sum game or a rescaled coordination game.*

Proof. By assumption we have that $b_{11} - b_{12} - b_{21} + b_{22}$ and $a_{11} - a_{12} - a_{21} + a_{22}$ be different from 0. Then, there exists $c \neq 0$ such that $b_{11} - b_{12} - b_{21} + b_{22} = c(a_{11} - a_{12} - a_{21} + a_{22})$. Let $D := cA - B$ then $d_{11} - d_{12} - d_{21} + d_{22} = 0$. Thus, for all i, j, we have $d_{ij} = d_{i2} + (d_{2j} - d_{22}) = u_i + v_j$, with $u_i := d_{i2}$ and $v_j := d_{2j} - d_{22}$. Since $ca_{ij} - b_{ij} = u_i + v_j$, we can define $c_{ij} = \frac{b_{ij} + u_i}{c} = a_{ij} - \frac{1}{c} v_j$. Hence, $(A, B) \sim (C, cC)$. $\square$

In case (I) it suffices to examine only rescaled zero-sum games, as best-response dynamics cannot cycle in rescaled coordination games given they are simply weighted potential games.[7] Similarly, since in case (III) the game has three isolated equilibria, it cannot be equivalent to a rescaled zero-sum game since such games have the same equilibrium set as some zero-sum game and thus their set of equilibria must form a convex polytope[8]. Thus, in case (III), it suffices to explore only rescaled coordination games.

---

7. In a rescaled coordination game, the ratio of the utilities of the agents is a constant that is independent of the action profile. Hence, any payoff-improving best-response for one agent is also payoff-improving for the other implying that any best response path must terminate at a Nash equilibrium.

8. Any rescaled zero-sum game has the same Nash equilibrium set as some zero-sum game. The set of Nash equilibria in each zero-sum game is a convex polytope (von Neumann and Morgenstern, 1944).

4.2 Case I: No pure Nash equilibrium (Rescaled Zero-Sum Games)

We will start off our analysis with Case I, concerning rescaled zero-sum games without a pure Nash equilibrium (and therefore having a unique interior Nash equilibrium). The analysis of strong swap regret minimization for RD will involve three steps. Firstly, we will show that all interior trajectories in this case are cycles. This topological argument hinges upon recent topological characterizations of RD trajectories in zero-sum games and variants thereof (Boone and Piliouras, 2019; Mertikopoulos et al., 2018; Nagarajan et al., 2020; Piliouras and Shamma, 2014; Piliouras et al., 2014). These works typically argue for weaker notions of divergence (e.g., Poincaré recurrence) for almost all initial conditions, whereas in our case we will need to prove periodicity of all interior trajectories. The closest result to our own is in Boone and Piliouras (2019), where periodicity of RD trajectories is argued for a different class of single-agent, evolutionary zero-sum games with three strategies. Our proof strategy will tailor those arguments to address our different class of games. Armed with this topological result, we will next prove a strong time-average property of 2×2 games with a unique interior Nash equilibrium and periodic RD trajectories. Specifically, we will show that along any such trajectory if we fix a specific mixed strategy of one of the agents and compute the time-average of their opponent conditioned only on those points where the first player applies the specified mixed strategy then the time-average of second agent will converge to the mixed Nash equilibrium. This is a significant strengthening of a well-known result that specifies that the time-average of external-regret minimization algorithms in zero-sum games *over their whole trajectory* converges to the marginal of a Nash equilibrium. Leveraging this strong property, it will easily follow that no agent can profitably deviate from their current play even if we allow them to use deviations which are conditional to their current mixed strategy, implying strong swap regret minimization in this case.

4.2.1 Periodicity of Interior Orbits

We will start of with a well-known theorem that constrains the possible limit behavior of smooth dynamical systems on a plane. Particularly, the Poincaré-Bendixson theorem provides sufficient conditions under which the limit behavior of such systems is periodic. Let $z = (x_1, y_1)$ denote the state of RD in a 2×2 rescaled zero-sum game where we remove the constrained variables x_2, y_2 to express that the system has two degrees of freedom. Formally, a limit set $\omega(z)$ of an initial condition z is the union of the limits of all its convergent subsequences, i.e., all its possible limit behaviors. Specifically, let $\phi : [0,1]^2 \times \mathbb{R} \to [0,1]^2$ be the flow of RD such that for any point $z \in [0,1]^2$, $\phi(z, -)$ defines a function of time corresponding to the trajectory of x. The ω-*limit (set) of z* is formally defined as the set of points $w \in [0,1]^2$ such that there exists a sequence (t_n) diverging to $+\infty$ such that $\phi(z, t_n) \to w$.

Theorem 4.4 (Poincaré-Bendixson theorem). *A limit set of a C^1 dynamical system over the plane, if non-empty and compact, that does not contain a rest point is a periodic orbit.*

We will apply the Poincaré-Bendixson theorem to show that all interior RD trajectories are periodic for these games.

Theorem 4.5. *Given any rescaled zero-sum game with a unique interior Nash equilibrium, all interior trajectories are periodic.*

Proof. First, from Piliouras et al. (2014), we have that the weighted sum of the Kullback-Leibler (KL) divergences of each agent's current mixed strategy from their respective interior Nash equilibrium strategy is a constant of the motion, i.e., it remains invariant over time for all orbits. The weights are strictly positive numbers such that, after multiplying these numbers with the current agent utilities, the rescaled game is zero-sum. The KL divergence is equal to zero if and only if the respective probability distributions are equal to each other. Thus, the fact that for all non-trivial (i.e., non-equilibrium) orbits, the original weighted sum of KL divergences is positive implies that any such trajectory will stay bounded away from the interior equilibrium. The only other fixed points for RD lie on the boundary and correspond to pure strategy profiles, however, since the Nash equilibrium is interior the KL divergence becomes unbounded as we approach the boundary and thus due to the invariance and non-negativity of KL divergences any interior trajectory will stay bounded away from the boundary. Due to the Poincaré-Bendixson theorem, we have than any non-trivial orbit of RD will stay bounded away from all fixed points and thus the resulting limit set will be a periodic orbit. Finally, it remains to show that not only are the limit set of all trajectories cycles, but all trajectories themselves are cycles.

We know that for any interior initial condition z, $\omega(z)$ is a periodic orbit and thus a Jordan curve of the plane $[0, 1]^2$, so it is separated into two connected components: an interior I and an exterior E.

By assumption, we know that there exists an interior rest point of RD, the interior Nash equilibrium. Let us denote it $z^\star$. We claim that this $z^\star$ has to be a point of I. Let us assume not, i.e., $z^\star \in E$. Since $\omega(z)$ is non-trivial, its interior I is non-empty. Pick a point $a \in I$. Draw a semi-infinite ray starting from $z^\star \in E$ in the direction of $a \in I$. Since I is bounded, this ray will transit from E to I then I to E at least once. Hence, it crosses $\omega(z)$ at least two times. Let us name those points ω_1 then ω_2. The sum of KL divergences is a strictly convex function, which attains its global minimum at $z^\star$, so will strictly increase as we advance along the ray. Thus, the weighted sum of KL divergence at ω_2 will be strictly larger than its value at ω_1, which contradicts the fact that it is a continuous invariant of motion. Therefore, we have that $z^\star \in I$; see the left panel of Fig. 3 for an illustration of the points constructed in this argument.

Summarizing, $\omega(z)$ is a cycle, and its time-average (the interior Nash equilibrium $z^\star$) lies in the interior I of $\omega(z)$. We now wish to show that the orbit starting from z is a periodic orbit. To prove that, we will argue that $z \in \omega(z)$. From $z^\star$, draw a semi-infinite line L in the direction of z. As I is bounded, this semi-infinite line must cross $\omega(z)$ in at least one point. Pick one of these points and call it z_ω. We claim that the weighted sum of KL divergences is equal to the sum of weighted KL divergences between $z^\star$ and z on L only at z_ω. Indeed, the sum of KL divergences is a strictly convex function with global minimum at $z^\star$ (equal to zero), so it is strictly increasing as one moves along the line L. Thus, z_ω is the unique intersection point between L and the level set of the weighted KL divergences with value equal to the weighted KL divergence between $z^\star$ and z. Therefore, we derive that $z = z_\omega \in \omega(z)$; see the right panel of Fig. 3 for an illustration of the points constructed in this argument. $\qquad\square$
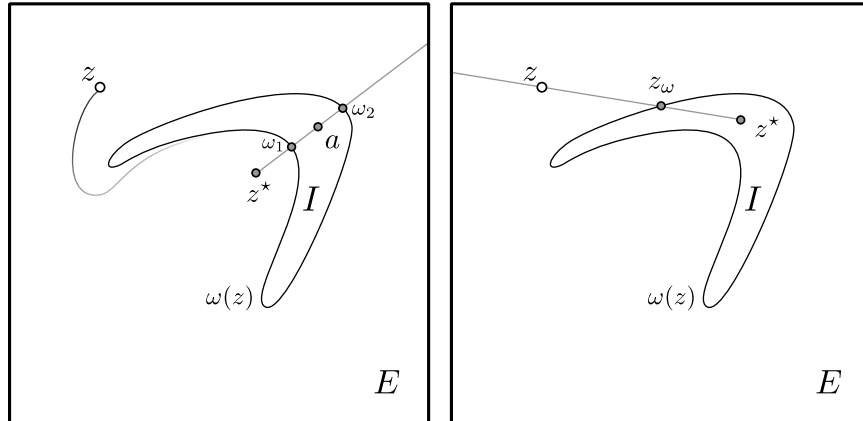
Figure 3: Illustrations of two key steps in the proof of Theorem 4.5. Left: A proof by contradiction that the unique Nash equilibrium is contained in the interior I of $\omega(z)$, the limit set of replicator dynamics initialised at z. If the Nash equilibrium z^* lies in the exterior E, then the points ω_1 and ω_2 constructed must have different weighted sums of KL divergences from z^*; a contradiction since they lie on the same level set for this function. Right: A proof by contradiction that $z \in \omega(z)$. If not, we construct $z_\omega \in \omega(z)$ which has smaller weighted sum of KL divergences from z^* than z. However, since this weighted sum of KL divergences is constant on trajectories of the dynamics, this is a contradiction unless $z = z_\omega$.

Next, we will move forward to prove strong time-average properties for periodic trajectories of RD dynamics in 2×2 games with unique interior Nash equilibrium, which naturally apply to our case of rescaled zero-sum games with unique interior Nash equilibrium.

4.2.2 Average Opponent Strategies

We first consider the average opponent strategy that an RD learner observes whenever playing a particular mixed strategy, a concept used to establish our subsequent regret-minimization proofs. In order to better understand the notion of 'time-average opponent strategy' derived here, consider the example in Fig. 4. Here, the left panel illustrates RD trajectories for a two-player 2-strategy game. The opponent strategy that the row player observes whenever their own strategy, x, is fixed to a particular value, σ, is illustrated via the highlighted vertical band in the figure; for a fixed orbit, whenever the row player passes through the mixed strategy $x = \sigma$, they observe one of two possible strategies for the (opposing) column player. The right panel illustrates the time-average of these observed opponent strategies, for which we seek to derive an analytic formula here.

Below, we derive the time-average opponent strategy result specifically for periodic 2×2 games (i.e., those wherein the dynamics cycle around an interior Nash equilibrium), as strong swap regret minimization can be handled more straightforwardly for other instances.

Proposition 4.6. *Consider a periodic 2×2 game with unique interior Nash equilibrium, i.e., a 2×2 game such that all interior orbits of RD are periodic. Let $(x^t, y^t)_{t \geq 0}$ be the RD trajectories with some fixed interior initial condition $(x^0, y^0) \in [0,1]^2$. Let $\sigma \in [0,1]$ denote*
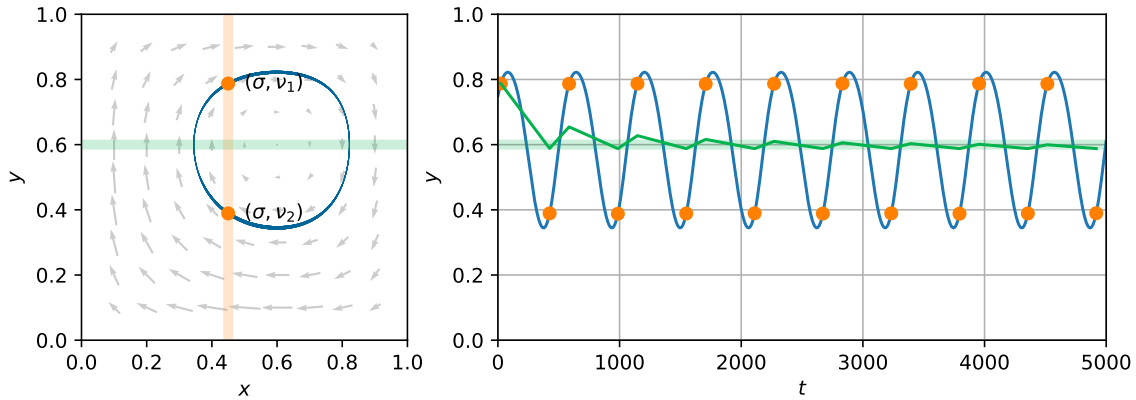
Figure 4: Illustration of RD for a two-player 2-strategy game. The left panel illustrates the player dynamics (with a particular periodic trajectory traced in blue and the corresponding line thickness indicating the row player's update speed, v_x). For a fixed orbit, whenever the row player passes through mixed strategy $\sigma \in [0,1]$, it observes one of two possible mixed strategies for the column player: $\nu_1, \nu_2 \in [0,1]$. The right panel illustrates the recurrence of these orbital crossings, with green series indicating the time-average convergence of the observed opponent strategies to their Nash distribution.

a fixed mixed strategy for the row player, and let $\nu_1, \nu_2 \in [0,1]$ denote the (at most) two possible mixed strategies played by the column player, i.e., such that both (σ, ν_1) and (σ, ν_2) lie on the orbit of RD.[9] Let (v_x^i, v_y^i) denote the gradient of RD at point (σ, ν_i), for $i = 1, 2$. Then the time-average opponent strategy observed by the row player up to time t is given by

$$\frac{\frac{1}{v_x^1}\nu_1 + \frac{1}{v_x^2}\nu_2}{\frac{1}{v_x^1} + \frac{1}{v_x^2}} + O\left(\frac{1}{t}\right). \tag{9}$$

Proof. We derive the time-average opponent strategy observed within time spans that are multiples of the period of the dynamics, T. It is sufficient to consider this case, as other cases can be dealt with by noting they modify the time-average observed strategy by at most a constant.[10]

Let $\varepsilon > 0$. The average strategy played by the column player whilst the row player plays a strategy within ε of σ is

$$\frac{\int_0^T y^t \mathbb{1}[x^t \in [\sigma - \varepsilon, \sigma + \varepsilon]] \, dt}{\int_0^T \mathbb{1}[x^t \in [\sigma - \varepsilon, \sigma + \varepsilon]] \, dt}. \tag{10}$$

---

9. On top of periodicity of all interior orbits, this proposition assumes that given any such orbit and for any fixed mixed strategy of the first (resp., second) player, the points of the orbit where the first (resp., second) agent applies that mixed strategy correspond to at most two mixed strategies of the second player. In Section 4.2.1, we have shown that the necessary conditions for this theorem are satisfied by all rescaled zero-sum games with a unique interior Nash equilibrium.

10. Specifically, writing $t = NT + r$, with $N \in \mathbb{N}_0$ and $r \in [0, T)$, the contribution of opponent strategies from the interval $(t - r, t)$ to the time-average up to time t goes as $O(\frac{1}{t})$.

We aim to derive a limiting expression as $\varepsilon \to 0$. Let $t_1, t_2 \in (0, T)$ be such that $(x^{t_1}, y^{t_1}) = (\sigma, \nu_1)$ and $(x^{t_2}, y^{t_2}) = (\sigma, \nu_2)$. Since the path of RD is differentiable, we have

$$x^{t_1+t} = \sigma + tv_x^1 + O(t), \quad y^{t_1+t} = \nu_1 + tv_y^1 + O(t)$$
$$x^{t_2+t} = \sigma + tv_x^2 + O(t), \quad y^{t_2+t} = \nu_2 + tv_y^2 + O(t),$$

where (v_x^i, v_y^i) denotes the gradient of RD at (σ, ν_i), for $i = 1, 2$. Then note that the condition $x^t \in [\sigma - \varepsilon, \sigma + \varepsilon]$ is satisfied for $t \in [t_1 - \frac{\varepsilon}{v_x^1} + O(\varepsilon), t_1 + \frac{\varepsilon}{v_x^1} + O(\varepsilon)] \cup [t_2 - \frac{\varepsilon}{v_x^2} + O(\varepsilon), t_2 + \frac{\varepsilon}{v_x^2} + O(\varepsilon)]$. Hence, the denominator integral in (10) simplifies to

$$2\frac{\varepsilon}{v_x^1} + 2\frac{\varepsilon}{v_x^2} + O(\varepsilon).$$

As for the numerator in (10), observe that

$$\int_0^T y^t \mathbb{1}[x^t \in [\sigma - \varepsilon, \sigma + \varepsilon]] \, dt = \sum_{i=1}^2 \int_{t_i - \frac{\varepsilon}{v_x^i} + O(\varepsilon)}^{t_i + \frac{\varepsilon}{v_x^i} + O(\varepsilon)} y^t \, dt$$

$$= \sum_{i=1}^2 \int_{-\frac{\varepsilon}{v_x^i} + O(\varepsilon)}^{\frac{\varepsilon}{v_x^i} + O(\varepsilon)} \left(\nu_i + tv_y^i + O(t) \right) \, dt$$

$$= \sum_{i=1}^2 \left(\left(2\frac{\varepsilon}{v_x^i} + O(\varepsilon) \right) \nu_i + O(\varepsilon) \right).$$

Altogether, the average strategy faced by the row player in (10) is rewritten as

$$\frac{\sum_{i=1}^2 \left(2\frac{\varepsilon}{v_x^i} + O(\varepsilon) \right) \nu_i + O(\varepsilon)}{2\frac{\varepsilon}{v_x^1} + 2\frac{\varepsilon}{v_x^2} + O(\varepsilon)}.$$

Letting $\varepsilon \to 0$, we compute

$$\lim_{\varepsilon \to 0} \frac{\sum_{i=1}^2 \left(\left(2\frac{\varepsilon}{v_x^i} + O(\varepsilon) \right) \nu_i + O(\varepsilon) \right)}{2\frac{\varepsilon}{v_x^1} + 2\frac{\varepsilon}{v_x^2} + O(\varepsilon)} = \lim_{\varepsilon \to 0} \left[\frac{(2\frac{\varepsilon}{v_x^1}\nu_1 + 2\frac{\varepsilon}{v_x^2}\nu_2)}{2\frac{\varepsilon}{v_x^1} + 2\frac{\varepsilon}{v_x^2} + O(\varepsilon)} + \frac{O(\varepsilon)}{2\frac{\varepsilon}{v_x^1} + 2\frac{\varepsilon}{v_x^2} + O(\varepsilon)} \right]$$

$$= \frac{\frac{1}{v_x^1}\nu_1 + \frac{1}{v_x^2}\nu_2}{\frac{1}{v_x^1} + \frac{1}{v_x^2}}.$$

$\square$

Remark 4.7. *In the representation* (9) *of the time-average opponent strategy induced by Proposition 4.6, the weights do not depend on the underlying parametrization choice. Namely, consider any reparameterization of mixed strategy* (x, y) *via* $(\theta^1, \theta^2) = (g(x), g(y))$, *where g is a non-decreasing differentiable function. Then, a similar line of arguments leads to the following result:*

$$\frac{\frac{1}{v_\theta^1} g(\nu_1) + \frac{1}{v_\theta^2} g(\nu_2)}{\frac{1}{v_\theta^1} + \frac{1}{v_\theta^2}} + O\left(\frac{1}{t} \right).$$

Proposition 4.6 implies in particular that, whenever the 2×2 game of interest is also zero-sum while exhibiting a unique interior Nash equilibrium, the time-average opponent strategy identifies to the Nash equilibrium itself. This result is proven in the following proposition.

Proposition 4.8. *Consider a periodic 2×2 game with unique interior Nash equilibrium, i.e., a 2×2 game such that all interior orbits of RD are periodic. Furthermore, assume that given any such periodic orbit and for any fixed mixed strategy of the first (resp., second) player, the points of the orbit where the first (resp., second) agent applies that mixed strategy correspond to at most two mixed strategies of the second player such that the equilibrium strategy of the second (resp., first) player lies in their convex combination.[11] Then, for any fixed first player strategy, the corresponding time-average opponent strategy in RD is the Nash strategy of the opponent.*

Proof. Consider the logit parameterization of strategies in these 2-strategy games, wherein a player's strategy is described by a single real number $\tau \in \mathbb{R}$, and the corresponding strategy probabilities are given by $e^{\tau}(1 + e^{\tau})^{-1}$ and $(1 + e^{\tau})^{-1}$. Given two joint strategies $(\sigma, \tau_1), (\sigma, \tau_2)$ in logit parameterization on the orbit of RD, the gradient for the row player under RD is given by the difference in payoffs for row strategies 1 and 2 against the column player's joint strategy, namely:

$$\begin{pmatrix} 1 & -1 \end{pmatrix} A \begin{pmatrix} e^{\tau_2}(1 + e^{\tau_2})^{-1} \\ (1 + e^{\tau_2})^{-1} \end{pmatrix},$$

and similarly for the column player. For shorthand, denote $z := (1, \ -1)$ and $y(\tau) := (e^{\tau}(1 + e^{\tau})^{-1}, \ (1 + e^{\tau})^{-1})^{\top}$. Per Proposition 4.6 and Remark 4.7, the average opponent strategy that the row player experiences when playing σ is expressed by

$$\frac{\frac{e^{\tau_1}(1+e^{\tau_1})^{-1}}{|zAy(\tau_1)|} + \frac{e^{\tau_2}(1+e^{\tau_2})^{-1}}{|zAy(\tau_2)|}}{\frac{1}{|zAy(\tau_1)|} + \frac{1}{|zAy(\tau_2)|}}.$$

Let $p_i := e^{\tau_i}(1 + e^{\tau_i})^{-1}$ so that $y(\tau_i) = (p_i, 1 - p_i)^{\top}$, for $i = 1, 2$. Similarly, let p_* denote the probability of the first strategy in the unique Nash equilibrium of the game, so that $y_* := (p_*, 1 - p_*)^{\top}$. For the statement of the proposition to be valid, it remains to prove the following relation:

$$\frac{\frac{1}{|zAy(\tau_1)|}p_1 + \frac{1}{|zAy(\tau_2)|}p_2}{\frac{1}{|zAy(\tau_1)|} + \frac{1}{|zAy(\tau_2)|}} = p_*,$$

which rewrites equivalently as

$$\frac{1}{|zAy(\tau_1)|}(p_1 - p_*) + \frac{1}{|zAy(\tau_2)|}(p_2 - p_*) = 0. \tag{11}$$

---

11. In Section 4.2.1, we have shown that the necessary conditions for this theorem are satisfied by all rescaled zero-sum games with a unique interior Nash equilibrium.

By definition of the Nash equilibrium y_*, recall that $zAy_* = 0$, so that

$$|zAy(\tau_i)| = |zAy(\tau_i) - zAy_*| = |zAz^\top| \times |p_i - p^*|, \quad \text{for } i = 1, 2 .$$

This allows one to rewrite the targeted relation (11) as

$$\frac{1}{|zAz^\top|} \left(\frac{(p_1 - p_*)}{|p_1 - p_*|} + \frac{(p_2 - p_*)}{|p_2 - p_*|} \right) = 0 ,$$

which holds true since $p_* \in [p_1, p_2]$ as (σ, τ_1) and (σ, τ_2) belong to the same orbit of RD. $\square$

We next prove the result establishing minimization of strong swap regret for Case I.

Theorem 4.9. *In any generic 2×2 game with a unique interior Nash equilibrium, RD minimizes strong swap regret.*

Proof. By Lemma 4.3 along with the fact that the game does not have any pure Nash equilibrium, we have that the game has to be equivalent to a rescaled zero-sum game. In such a case, we have that all interior trajectories of RD are periodic. Following the line of arguments of Proposition 4.6 and Proposition 4.8, we can conclude that RD minimizes strong swap regret, since for any fixed first player strategy, the average opponent strategy it observes is the Nash equilibrium component of the opponent and thus no profitable deviation is possible even if the agent is allowed to condition their deviating strategy to their current mixed strategy. $\square$

The argument about the time-average convergence to the Nash equilibrium is specific to zero-sum games with a unique interior Nash equilibrium. In some sense, this is the most interesting subclass in which to study strong swap regret minimization, since this minimization result is itself strongest in this class of games; we will see below that in other cases, strong swap regret minimization coincides with weaker notions of regret.

4.3 Case II: Games with a Unique Pure Nash Equilibrium

In the case with a unique Nash equilibrium on the simplex boundary, as argued above, RD converges to the pure Nash equilibrium, which we will exploit to establish a link between external regret and strong swap regret.

Theorem 4.10. *In any generic 2×2 game with a unique pure Nash equilibrium, RD minimizes strong swap regret.*

Proof. We will first argue that in this case the game is strictly dominance solvable. Namely, after the iterated elimination of pure strategies that are strictly dominated by other pure strategies, we are left with a single strategy for each agent. By our genericity assumption we have that for each pure strategy of their opponent, each player has a unique (pure) best response. Since the best response dynamic has a unique sink (the unique pure/strict Nash equilibrium) for at least one of the two agents one strategy strictly dominates the other. Once that dominated strategy has been eliminated, the genericity of the game will lead to the elimination of one of the two remaining strategies of the other agent, implying that the game is strictly dominance solvable.

In such games, RD (as well as a wide range of other evolutionary dynamics) is bound to converge to playing the (iteratively) undominated strategies with probability one (Weibull, 1997, Proposition 4.6). However, in any small enough neighborhood around this strict Nash equilibrium, the best response of each agent is unique and and it is their strategy at the unique pure/strict Nash equilibrium. Hence, given any interior initial condition, after a finite amount of time, the optimal deviating strategies for strong swap regret are going to be independent of the (exact) mixed strategy of the agent and, instead, they will stay fixed and equal to the Nash equilibrium strategy of that agent. Thus, strong swap regret will be within a constant of the external regret of RD, which is known to be bounded in general games. Thus, RD also minimizes strong swap regret. □

4.4 Case III: Games with Two Pure Nash Equilibria (Rescaled Coordination Games)

This is the last case of 2×2 games, where we examine games with multiple Nash equilibria. Given the fact that some trajectories converge to pure Nash equilibria and some to mixed Nash equilibria, we need specialized arguments.

Lemma 4.11. *In any 2×2 games with several Nash equilibria on the simplex boundary, RD minimizes strong swap regret.*

Proof. In this case, the game is equivalent to a rescaled coordination game per Section 4.1. However, rescaled coordination games are weighted potential games where the potential function of the player is the utility of, e.g., the first player. In this case, RD is known to converge to Nash equilibria for all interior initial conditions (Papadimitriou and Piliouras, 2018). In terms of the possible limit points of different trajectories, only two situations can occur: either the limit is a pure Nash equilibrium, or it is an interior (mixed) Nash equilibrium.

In the case where the trajectory converges to a pure Nash equilibrium, the boundedness of strong swap regret follows from the same argument as in the case of strictly dominance solvable games converging to a pure Nash equilibrium, see the proof of Theorem 4.10. Thus, the only remaining case is that of trajectories converging to a mixed Nash equilibrium.

In the case of rescaled coordination games with interior Nash equilibrium, there always exists a positive weight such that the weighted difference of KL divergences between each agent's mixed Nash equilibrium and their evolving state remains invariant over time (see Hofbauer 1996; Nagarajan et al. 2018). Specifically, given any pair of (interior) mixed strategies $x = (x_1, x_2)$ and $y = (y_1, y_2)$, there exists $w > 0$ such that the KL divergence

$$w D_{\mathrm{KL}}\big(p \parallel x^t\big) - D_{\mathrm{KL}}\big(q \parallel y^t\big) = w \sum_i p_i \ln(\frac{p_i}{x_i^t}) - \sum_j q_i \ln(\frac{q_j}{y_j^t})$$

between the interior Nash equilibrium (p, q) and the state (x^t, y^t) of the system is invariant over time. Clearly, the value of this invariant function evaluated at the mixed Nash equilibrium (p, q) is equal to zero and by continuity the value of this function must be equal to zero for any trajectory that converges to that mixed Nash equilibrium. The important property implied by this invariant function is that any trajectory that converges to the mixed Nash equilibrium (p, q) cannot intersect the lines $x_1 = p_1$, $y_1 = q_1$. If that was the case and

exactly one of the two conditions $x_1 = p_1$, $y_1 = q_1$ was satisfied, then exactly one of the two KL divergence terms would be equal to zero, implying that the invariant function cannot be equal to zero. On the other hand, if the trajectory intersects the lines $x_1 = p_1$, $y_1 = q_1$ simultaneously then by uniqueness of the solution of the ODE this is the trivial equilibrium trajectory that does not converge to equilibrium since it already starts at it. Hence, when encoding the state of the 2×2 replicator dynamics in the $[0,1]^2$ plane expressing the variables x_1, y_1, then any trajectory converging to the interior Nash equilibrium must be strictly contained in one of the four parallelograms encoded by the lines $x_1 = p_1$, $y_1 = q_1$. Thus, the best response of both agents along any such trajectory remains fixed over time and, as a result, the maximal strong swap regret is achieved by deviating to that fixed best response throughout the system orbit. Therefore, strong swap regret is upper bounded by standard external regret, which is well-known to be bounded for RD (e.g., see Mertikopoulos et al. 2018) and the proof is complete. □

5. Linking Strong Swap Regret to Other Optimization Notions

Having established that RD minimizes strong swap regret in generic 2×2 games, we now present relationships to more standard perspectives in online learning.

5.1 Strong Swap Regret versus External Regret Minimization

A natural question to ask is whether minimization of strong swap regret follows immediately from minimization of standard (external) regret in the special case of 2×2 games. The following theorem shows that this is not the case.

Theorem 5.1. *There exist algorithms that minimize external regret, which exhibit unbounded strong swap regret in some 2×2 games.*

Proof. Consider an arbitrary tuple of algorithms that minimize external regret (e.g., bounded external regret algorithms such as RD), with one such algorithm used by each of the agents. Given such a tuple of algorithms, we define a slight variant of them that initializes them as playing a pre-computed CCE in a chosen target game that does not have bounded strong swap regret. For example, consider the 2×2 game where both agents receive a payoff of 1 if they choose the same strategy (i.e., both play strategy a or b), otherwise receiving a payoff of 0. Now, consider the following strategy distribution in this game: with probability $1/2$, both agents play the mixed strategy where they choose the first strategy with probability $3/4$; with probability $1/2$, both agents play the mixed strategy where they choose the second strategy with probability $3/4$. This overall strategy distribution is a CCE.[12] Thus, dividing the time interval into epochs of length one, we can have the agents reproduce this CCE by alternating between playing the two described mixed strategies.

   The agents continue producing this predefined play as long as no deviation from it is observed. If any deviation is observed from any individual agent, then all agents subsequently deviate from this play, from this point onward always applying their assigned regret-minimizing algorithm (e.g., RD). Clearly, each of these algorithms still minimizes

---

12. The overall distribution chooses pure strategy profiles $(a, a), (b, b)$ each with probability $5/16$ and the profiles $(a, b), (b, a)$ with probability $3/16$; it is easy to verify this distribution satisfies the CCE definition.

external regret, as even if a deviation is observed leading to this switch away from the original CCE mixed strategy, the total regret experienced on any single period is bounded and the total regret of the individual algorithm used (e.g., RD) is bounded as well. On the other hand, by construction the resulting play does not minimize strong swap regret, as any of the two mixed strategies applied by the agents is not a best response during the time epoch that is being played; thus, instead the agents would rather deviate from playing their assigned strategy, not with probability 3/4, but with probability 1. □

5.2 Strong Swap Regret in the Case of Online Learning With Two Actions

While the definition of strong swap regret was motivated by the need for more granular techniques in the study of learning in games, a related question is whether this notion also yields insights for more traditional online learning settings. Below, we show that it is possible to provide positive regret results even if we allow for the deviating strategies to depend on the current mixed strategy, the telltale attribute of strong swap regret, under specific conditions on the learning trajectory. Specifically, we explore the conditions under which we can prove positive results for strong swap regret, not in the case of 2×2 games, but rather in the case of online learning with two actions (i.e., strategies) whose utilities evolve continuously over time.

Theorem 5.2. *Given bounded and continuous utility functions u_1^t, u_2^t, the cumulative regret of RD remains bounded even if we consider only the part of the history of play where replicator chooses the first action with probability $p \in (p_0 - \epsilon, p_0 + \epsilon)$ for any $p_0 \in (0, 1)$, as long as ϵ is small enough such that the replicator trajectory always exits the interval from the opposite end of the interval from the one it enters it.*

Proof. Let us create a possibly infinite sequence of time intervals (a_j, b_j) whose union captures the set of time instances t such that $p^t \in (p_0 - \epsilon, p_0 + \epsilon)$. To explore the total regret for a player in online learning in the time intervals where the probability of playing their first strategy is between $p_0 - \epsilon$ and $p_0 + \epsilon$, it suffices to track the total regret for deviating against any fixed strategy for the union of (a_j, b_j) intervals.

In the case of RD in any such interval (a_j, b_j), the regret for not deviating to some fixed strategy i is given as follows:

$$\dot{p}_i = p_i \left(u_i - \sum_i p_i u_i \right).$$

Rearranging,

$$\frac{\dot{p}_i}{p_i} = u_i - \sum_i p_i u_i$$

$$\therefore \ln(p_i(b_j)) - \ln(p_i(a_j)) = \int_{a_j}^{b_j} u_i^\tau d\tau - \int_{a_j}^{b_j} \sum_i p_i^\tau u_i^\tau d\tau \,.$$

Without loss of generality, we can assume that the trajectory enters the $(p_0 - \epsilon, p_0 + \epsilon)$ interval from below. In this case, the regret in the union of these times intervals is

$$\sum_j \ln(p_i(b_j)) - \ln(p_i(a_j)) = \ln(p_0 + \epsilon) - \ln(p_0 - \epsilon) + \ln(p_0 - \epsilon) - \ln(p_0 + \epsilon) + \dots$$

$$\leq \ln(p_0 + \epsilon) - \ln(p_0 - \epsilon)$$

and thus remains bounded for all time. □

Overall, these insights indicate that there exist interesting positive and negative results linking strong swap regret to classical notions of regret.

6. Experiments

In this section, we ground our theoretical findings in experimental analysis of RD in numerous games of varying characteristics.

6.1 Φ-Regret Minimization via RD in 2×2 Games

As mentioned earlier, practical experiments preclude one from demonstrating that strong swap regret is minimized, as time-discretization implies that any individual mixed strategy observed in a single RD trajectory may not be revisited in future trajectories. Thus, we focus our empirical analysis here on mosaic regret, the practically-realizable instance of Φ-regret we introduced in Section 3.3.

We first conduct a set of experiments illustrating the minimization of mosaic regret under RD in a wide suite of 2×2 games. As noted earlier, despite their simple structure, these payoffs form an important class capturing numerous canonical games that have received tremendous attention throughout the game theory literature (e.g., Matching Pennies, Prisoner's Dilemma, Stag Hunt, etc.) (Bruns, 2015; Guyer and Rapoport, 1972; Klos et al., 2010; Rapoport, 1966; Robinson and Goforth, 2005).

The specific games we consider are those defined by Bruns (2015), which taxonomizes a collection of 2×2 games into several distinct classes by considering the patterns of payoffs received by each player. Bruns (2015) identifies 12 sets of basis payoffs corresponding to canonical games of varying characteristics, summarized for the row player in Fig. 5(a); corresponding column player payoffs for these games are defined in Bruns (2015) as the transpose along the *anti*-diagonal, which we also use for consistency.

The combination of these basis payoffs enables the definition of a large collection of 144 2×2 games with varying characteristics (e.g., cyclical games, win-win games, pure coordination games, etc.). For each of these games, we visualize the vector field summarizing the behavior of agents playing RD in Fig. 5(b). The x- and y-axes here, respectively, show the row and column player's probabilities of playing their first strategies. Figure 6 visualizes an evaluation of the mosaic regret across this diverse set of games. For simplicity, the partitioning scheme Σ we use for computing mosaic regret is to subdivide the first player's mixed strategy space x into 10 discrete, equally-sized bins. To generate the mosaic regret results, for each game we run 10 trials of RD (each with an independently-initialized set of seed strategies for the row and column players). We plot the time-average mosaic

| Chicken (Ch)	Battle (Ba)	Hero (Hr)	Compromise (Cm)	Deadlock (Dl)	Prisoner's dilemma (Pd)	Stag hunt (Sh)	Assurance (As)	Coordination (Co)	Peace (Pc)	Harmony (Ha)	Concord (Nc)
$\begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 \\ 1 & 4 \end{bmatrix}$	$\begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix}$	$\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}$	$\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}$	$\begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix}$	$\begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix}$	$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$

(a)

(b)
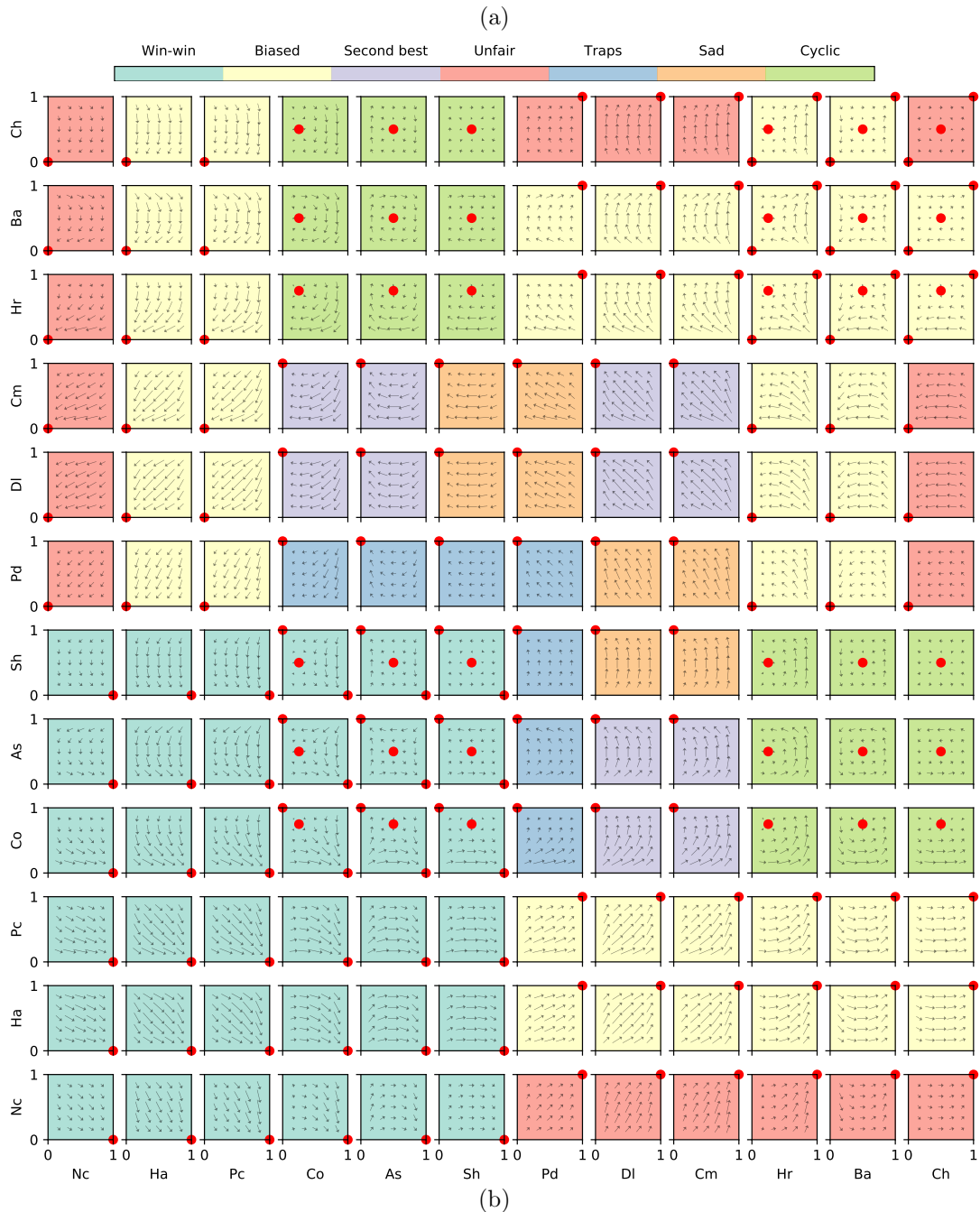
Figure 5: (a) Row player payoffs for 2×2 games defined by Bruns (2015). Note: in the convention of Bruns (2015), corresponding column player payoffs are defined as a transpose along the anti-diagonal. (b) RD run on 144 games composed of the indicated combinations of row and column player payoffs. Each subplot shows the dynamics for RD on a distinct game, with background colors corresponding to the classes of games originally defined by Bruns (2015). Nash equilibria for each game are indicated by the red points.
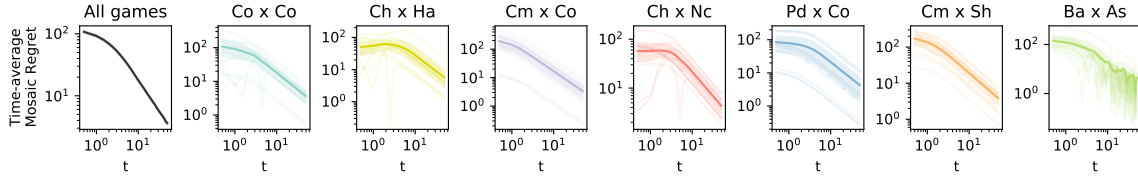
Figure 6: Mosaic regret convergence for 2×2 games. For each of the 144 games (shown earlier in Fig. 5(b)), we run 10 trials of RD with independently random initial strategies for both the row and column players, computing mosaic regret through time. The leftmost subplot above summarizes the time-average convergence of mosaic regret, $1/T \sum_{t=0}^{T} MR^t$, across all games and trials. The subsequent subplots show convergence for one example game per Bruns class, with classes corresponding to those in Fig. 5(b); in each of these latter subplots, each line indicates an individual trial, and the shaded region shows the average mosaic regret and 95% confidence interval across all trials.

regret averaged across all 1440 combinations of games and trials in the lefternmost subplot of Fig. 6, verifying our earlier theoretical convergence results for these general classes of games. Moreover, to understand the game-specific convergence characteristics, we plot the time-average mosaic regret for one example game per Bruns class (i.e., one per 'win-win', 'biased', 'second best', 'unfair', 'traps', 'sad', and 'cyclic' class of games) in the subsequent subplots of Fig. 6. The time-average mosaic regret converges across all these games, with a notable observation being that final game, Ba × As, suffers from higher-variance across trials, which is explained primarily due to the cyclical nature of this particular game.

6.2 Mosaic Regret Beyond 2×2 Games

We next consider several increasingly-complex instances of larger games, where we observe evidence that RD minimizes mosaic regret. Each row of Fig. 8 visualizes results associated with a distinct 3×3 game, where the respective row player payoffs are defined as

$$A_1 = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \qquad A_2 = \begin{bmatrix} 0 & -1 & 2 \\ 1 & 0 & -1 \\ -2 & 1 & 0 \end{bmatrix} \qquad A_3 = \begin{bmatrix} 1 & -1 & 1.2 \\ 1 & 0 & -1 \\ -1 & 1 & -0.5 \end{bmatrix}, \qquad (12)$$

with corresponding column player payoffs $B_i = -A_i$ for each game i. These payoffs, in order, correspond to the canonical game of Rock–Paper–Scissors (RPS), a variant of RPS with symmetrically-biased payoffs (yielding an non-uniform Nash equilibrium for both players), and an asymmetric variant of RPS (yielding distinct Nash equilibria for either player).

We next construct a particular instance of a mosaic regret deviation partition, and illustrate that RD minimizes mosaic regret under such a partitioning scheme in the above games. The first column in Fig. 8 visualizes points corresponding to raw trajectories exhibited under RD. For each game, the second column of Fig. 8 visualizes a specific mosaic regret deviation partition as follows: we select an arbitrary reference point in the row player's trajectories, x_r, and compute the KL-divergence of the point with respect to the row player's Nash equilibrium, x_* (indicated in orange in the corresponding plots), i.e., $d \triangleq D_{\mathrm{KL}}(x_r \parallel x_*)$. Subsequently, throughout all raw trajectories, we retain only the joint
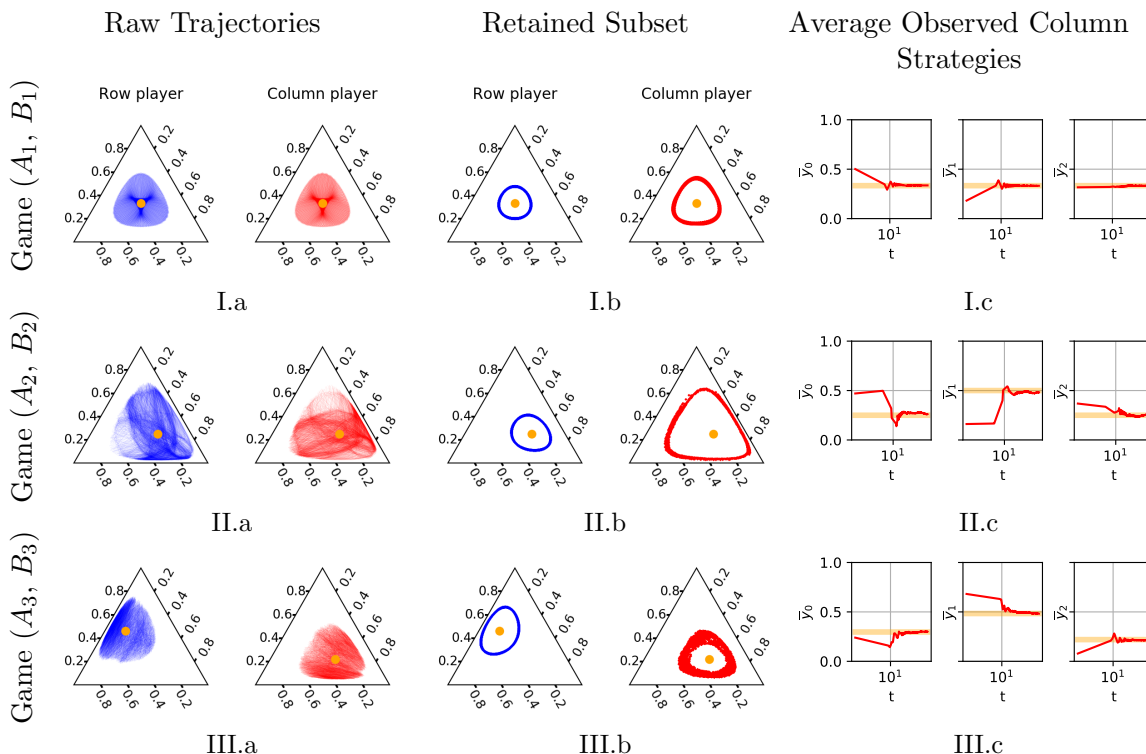
Figure 8: Several instances of 3×3 games where we observe evidence that RD minimizes mosaic regret. In all plots, the Nash equilibrium is indicated in orange for each player. Each row corresponds to a zero-sum game (with payoffs described in the main text). The first column visualizes points corresponding to raw trajectories exhibited under RD. Next, we retain only the partition of points wherein the first player's strategy, x, is within a small neighborhood of a reference KL-divergence from its Nash equilibrium; the second column plots the joint strategy under this filtering scheme. In the final column, we use all such sampled column player strategies up to the given timestep t, $y_{<t}$, to compute the time-average opponent strategy, $\bar{y}^t$. The notable observation is that the time-average opponent observed strategy converges to the column player's Nash equilibrium (in orange) in each instance, under the described partitioning scheme.

strategies (x, y) wherein the first player's distance to the Nash equilibrium is close to this reference KL-divergence (i.e., $D_{\mathrm{KL}}(x \parallel x_*) \in [d - \varepsilon, d + \varepsilon]$, where ε is a small threshold parameter). The set of points that constitute the resulting partition are visualized in the 'Retained Subset' column of Fig. 8. Next, as in Proposition 4.6, for each timestep t, we use all such sampled column player strategies up to that timestep, $y_{<t}$, to compute the time-average opponent strategy, $\bar{y}^t$. The third column of Fig. 8 visualizes this time-average 'observed' column strategy for each of the games, with the notable observation being that it converges to the column player's Nash equilibrium (in orange) for each game. In other words, as in our earlier 2×2 analysis and experiments, RD minimizes mosaic regret in these larger games, under the partitioning scheme described.

However, this property may not necessarily hold in more general settings. For example, consider again the game of Rock–Paper–Scissors as defined by A_1 in (12), but where both players' strategies are initialized identically (i.e., $x^0 = y^0$), thus ensuring that $x^t = y^t$ for all t under RD due to symmetry of the game payoffs. Subsequently, both players consistently attain the value of the game due to the symmetry in chosen strategies. In such a scenario, at any time t, either player can increase their payoff by best-responding, thus ensuring unbounded mosaic regret. Such counterexamples, thus, motivate a need to more thoroughly investigate strong swap regret and mosaic regret convergence in larger and more general classes of games.

7. Discussion

There has been increasing interest in recent years in defining alternative, generalized notions of regret to better enable stronger theoretical guarantees drawn about learning algorithms (Lehrer, 2003). For instance, Mohri and Yang (2014) introduce *conditional swap regret*, which generalizes swap regret to enable deviations to condition on the learner's history of actions (as opposed to the standard unconditioned deviations). Arora et al. (2012a) investigate strengthening of the utility classes (resp., 'adversary classes' in their case, as they focus on losses), enabling adaptation of the utilities (resp., losses defined by the adversary) on the learner's past actions. Morrill et al. (2021b) explore existing and defines new forms of strategy deviations in the space of extensive-form games. *Dynamic regret* (Zinkevich, 2003), measures the regret with respect to a sequence of changing baselines, and likewise *adaptive regret* (Daniely et al., 2015; Hazan and Seshadhri, 2007) computes regret by considering utility differences in windows of time intervals of a specific length. Overall, the primary motivation behind this line of investigation has been similar to ours: to generalize canonical forms of regret by introducing degrees of freedom in terms of how regret is computed (e.g., by permitting more general classes of deviations, varying the time intervals over which regret is computed, allowing the baseline compared against to change in various ways, etc.).

In our work, we illustrated through a re-examination of RD that, under the lens of Φ-regret, one can study stronger properties associated with the real-time behaviors of such learning processes in games, thus establishing strong convergence guarantees even over the space of mixed strategies. To our knowledge, our work provides a first demonstration of the strongest form of Φ-regret being minimized by the well-studied and simple-to-implement RD algorithm in both theory and practice. While our theoretical and empirical results focused primarily on exhaustively exploring in the space of generic 2×2 two-player zero-sum games, examination of larger games under this paradigm will make for an interesting avenue of future exploration. Moreover, another line of future work will involve investigating whether discrete-time dynamical systems related to RD, such as multiplicative weights update, also minimize strong swap or mosaic regret in similar or more general circumstances.

Acknowledgements

References

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *In Proceedings of the Twenty-Ninth International Conference on Machine Learning*, January 2012a.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012b.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 321–338, 2018.

James P. Bailey and Georgios Piliouras. Fast and furious learning in zero-sum games: Vanishing regret with non-vanishing step sizes. In *Advances in Neural Information Processing Systems*, volume 32, pages 12977–12987, 2019.

Bikramjit Banerjee and Jing Peng. Performance bounded reinforcement learning in strategic interactions. In *AAAI*, volume 4, pages 2–7, 2004.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Intell. Res. (JAIR)*, 53:659–697, 2015.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun):1307–1324, 2007a.

Avrim Blum and Yishay Mansour. Learning, regret minimization, and equilibria. In Noam Misan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, chapter 4. Cambridge University Press, 2007b.

Victor Boone and Georgios Piliouras. From Darwin to Poincaré and von Neumann: Recurrence and cycles in evolutionary and algorithmic game theory. In *International Conference on Web and Internet Economics*, pages 85–99. Springer, 2019.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17057–17069. Curran Associates, Inc., 2020.

Bryan Randolph Bruns. Names for games: locating 2× 2 games. *Games*, 6(4):495–520, 2015.

Andrea Celli, Alberto Marchesi, Gabriele Farina, and Nicola Gatti. No-regret learning dynamics for extensive-form correlated equilibrium. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Yu-Han Chang. No regrets about no-regret. *Artif. Intell.*, 171(7):434–439, 2007.

Yu-Han Chang and Leslie Pack Kaelbling. Hedged learning: Regret-minimization with learning experts. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 121–128. Association for Computing Machinery, 2005. ISBN 1595931805.

Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *COLT*, 2019.

Yun Kuen Cheung and Georgios Piliouras. Chaos, extremism and optimism: Volume analysis of learning in games. *Advances in Neural Information Processing Systems*, 33, 2020.

Yun Kuen Cheung and Georgios Piliouras. Online optimization in games via control theory: Connecting regret, passivity and poincaré recurrence. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1855–1865. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/cheung21a.html.

Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. The route to chaos in routing games: When is price of anarchy too optimistic? *Advances in Neural Information Processing Systems*, 33, 2020a.

Thiparat Chotibut, Fryderyk Falniowski, Michal Misiurewicz, and Georgios Piliouras. Family of chaotic maps from game theory. *Dynamical Systems*, 2020b. doi: 10.1080/14689367. 2020.1795624.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, pages 1405–1411. PMLR, 2015.

Gabriele Farina, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium. *arXiv preprint arXiv:2104.01520*, 2021a.

Gabriele Farina, Andrea Celli, and Tuomas Sandholm. Efficient decentralized learning dynamics for extensive-form coarse correlated equilibrium: No expensive computation of stationary distributions required. *arXiv preprint arXiv:2109.08138*, 2021b.

Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *NeurIPS*, 2020.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

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

Aram Galstyan. Continuous strategy replicator dynamics for multi-agent q-learning. *Auton. Agents Multi Agent Syst.*, 26(1):37–53, 2013.

Nicola Gatti and Marcello Restelli. Sequence-form and evolutionary dynamics: Realization equivalence to agent form and logit dynamics. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 509–515. AAAI Press, 2016.

Michel Goemans, Vahab Mirrokni, and Adrian Vetta. Sink equilibria and convergence. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 142–151. IEEE, 2005.

Geoffrey J Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *International Conference on Machine Learning (ICML)*, 2008.

Amy Greenwald and Amir Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Conference on Learning Theory (COLT)*, 2003.

Melvin J Guyer and Anatol Rapoport. 2× 2 games played once. *Journal of Conflict Resolution*, 16(3):409–431, 1972.

James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Christopher Harris. On the rate of convergence of continuous-time fictitious play. *Games and Economic Behavior*, 22(2):238–259, 1998.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. In *Electronic colloquium on computational complexity (ECCC)*, volume 14, 2007.

Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, et al. Neural replicator dynamics: Multiagent learning via hedging policy

gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 492–501, 2020.

Josef Hofbauer. Evolutionary dynamics for bimatrix games: A hamiltonian system? *Journal of mathematical biology*, 34(5):675–688, 1996.

Josef Hofbauer and Karl Sigmund. *Evolutionary games and population dynamics*. Cambridge university press, 1998.

Sue Ann Hong. A practical no-linear-regret algorithm for convex games. 2008.

Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *ACM Symposium on Theory of Computing (STOC)*, 2009.

Tomas Klos, Gerrit Jan Van Ahee, and Karl Tuyls. Evolutionary dynamics of regret minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 82–96. Springer, 2010.

Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics & Games*, 4(2):125, 2017.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Ehud Lehrer. A wide range no-regret theorem. *Games and Economic Behavior*, 42(1): 101–115, 2003.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

Mehryar Mohri and Scott Yang. Conditional swap regret and conditional correlated equilibrium. *Advances in Neural Information Processing Systems*, 27:1314–1322, 2014.

Barnabé Monnot and Georgios Piliouras. Limits and limitations of no-regret learning in games. *Knowledge Eng. Review*, 32:e21, 2017.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Dustin Morrill, Ryan D'Orazio, Marc Lanctot, James R Wright, Michael Bowling, and Amy R Greenwald. Efficient deviation types and learning for hindsight rationality in extensive-form games. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7818–7828. PMLR, 18–24 Jul 2021a. URL `https://proceedings.mlr.press/v139/morrill21a.html`.

Dustin Morrill, Ryan D'Orazio, Reca Sarfati, Marc Lanctot, James R Wright, Amy R Greenwald, and Michael Bowling. Hindsight and sequential rationality of correlated play. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5584–5594, May 2021b. URL https://ojs.aaai.org/index.php/AAAI/article/view/16702.

Hervé Moulin and J-P Vial. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3-4):201–221, 1978.

Sai Ganesh Nagarajan, Sameh Mohamed, and Georgios Piliouras. Three body problems in evolutionary game dynamics: Convergence, periodicity and limit cycles. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 685–693. International Foundation for Autonomous Agents and Multi-agent Systems, 2018.

Sai Ganesh Nagarajan, David Balduzzi, and Georgios Piliouras. From chaos to order: Symmetry and conservation laws in game dynamics. In *International Conference on Machine Learning*, pages 7186–7196. PMLR, 2020.

J. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, pages 48–49, 1950.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. α-rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9(1):9937, Jul 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-45619-9.

Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *Advances in Neural Information Processing Systems*, pages 5872–5882, 2017.

Christos Papadimitriou and Georgios Piliouras. From Nash equilibria to chain recurrent sets: Solution concepts and topology. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 227–235. ACM, 2016.

Christos Papadimitriou and Georgios Piliouras. From Nash equilibria to chain recurrent sets: An algorithmic solution concept for game theory. *Entropy*, 20(10):782, 2018.

Christos Papadimitriou and Georgios Piliouras. Game dynamics as the meaning of a game. *ACM SIGecom Exchanges*, 16(2):53–63, 2019.

Georgios Piliouras and Jeff S Shamma. Optimization despite chaos: Convex relaxations to complex limit sets via Poincaré recurrence. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 861–873. SIAM, 2014.

Georgios Piliouras, Carlos Nieto-Granda, Henrik I. Christensen, and Jeff S. Shamma. Persistent patterns: Multi-agent learning beyond equilibrium and utility. In *AAMAS*, pages 181–188, 2014.

Anatol Rapoport. A taxonomy of 2× 2 games. *General systems*, 11:203–214, 1966.

David Robinson and David Goforth. *The topology of the 2x2 games: a new periodic table*, volume 3. Psychology Press, 2005.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. In *Proc. of STOC*, pages 513–522, 2009.

Tim Roughgarden. *Twenty lectures on algorithmic game theory*. Cambridge University Press, 2016.

James B. T. Sanders, J. Doyne Farmer, and Tobias Galla. The prevalence of chaotic dynamics in games with many players. *Scientific Reports*, 8, 2018.

Tuomas Sandholm. The state of solving large incomplete-information games, and application to poker. *Ai Magazine*, 31(4):13–32, 2010a.

William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, 2010b.

Yuzuru Sato, Eizo Akiyama, and J Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.

Peter Schuster and Karl Sigmund. Replicator dynamics. *Journal of theoretical biology*, 100 (3):533–538, 1983.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

Stratis Skoulakis, Tanner Fiez, Ryann Sim, Georgios Piliouras, and Lillian Ratliff. Evolutionary game theory squared: Evolving agents in endogenously evolving zero-sum games. In *AAAI*, 2021.

Sylvain Sorin. Exponential weight algorithm in continuous time. *Mathematical Programming*, 116(1-2):513–528, 2009.

Sylvain Sorin. Replicator dynamics: Old and new. *Journal of Dynamics & Games*, 7(4): 365, 2020.

Sriram Srinivasan, Marc Lanctot, Vinícius Flores Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3426–3439, 2018.

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

Peter D Taylor and Leo B Jonker. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2):145–156, 1978.

Karl Tuyls and Simon Parsons. What evolutionary game theory tells us about multiagent learning. *Artif. Intell.*, 171(7):406–416, 2007.

Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *The Second International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2003, July 14-18, 2003, Melbourne, Victoria, Australia, Proceedings*, pages 693–700, 2003.

Karl Tuyls, Pieter Jan't Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12(1):115–153, 2006.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Jörgen W Weibull. *Evolutionary game theory*. MIT press, 1997.

Michael Wunder, Michael L. Littman, and Monica Babes. Classes of multiagent q-learning dynamics with epsilon-greedy exploration. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 1167–1174. Omnipress, 2010.

Hobart Peyton Young. *Strategic learning and its limits*. OUP Oxford, 2004.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, 2003.

Martin Zinkevich, Michael Johanson, Michael H Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *NIPS*, volume 7, page 1729, 2007.