Multi-Agent Learning for Iterative Dominance Elimination: Formal Barriers and New Algorithms

Jibang Wu[*1], **Haifeng Xu**[*1], **Fan Yao**[*1]

[1] Department of Computer Science, University of Virginia, USA
jw7jb@virginia.edu, hx4ad@virginia.edu, fy4bc@virginia.edu,

Abstract

Dominated actions are natural (and perhaps the simplest possible) multi-agent generalizations of *sub-optimal* actions as in standard single-agent decision making. Thus similar to standard bandit learning, a basic learning question in multi-agent systems is whether agents can learn to efficiently eliminate *all* dominated actions in an unknown game if they can only observe noisy bandit feedback about the payoff of their played actions. Surprisingly, despite a seemingly simple task, we show a quite negative result; that is, standard no regret algorithms — including the entire family of Dual Averaging algorithms — provably take *exponentially* many rounds to eliminate all dominated actions. Moreover, algorithms with the stronger no swap regret also suffer similar exponential inefficiency. To overcome these barriers, we develop a new algorithm that adjusts `Exp3` with `D`iminishing `H`istorical rewards (termed `Exp3-DH`); `Exp3-DH` gradually "forgets" history at carefully tailored rates. We prove that when all agents run `Exp3-DH` (a.k.a., *self-play* in multi-agent learning), all dominated actions can be iteratively eliminated within polynomially many rounds. Our experimental results further demonstrate the efficiency of `Exp3-DH`, and that state-of-the-art bandit algorithms, even those developed specifically for learning in games, fail to eliminate all dominated actions efficiently.

Introduction

Two seminal results in multi-agent learning are that agents using no regret learning algorithms will converge to a coarse correlated equilibrium (CCE) whereas the stronger no-swap regret learning algorithms will bring agents to a correlated equilibrium (CE) (Foster and Vohra 1999; Blum and Mansour 2005). Crucially, in both results, the converging sequence is the *average* of agents' historical plays. Recent results show that this does not imply the convergence of agents' final actions (Mertikopoulos, Papadimitriou, and Piliouras 2018; Daskalakis and Panageas 2018; Lin, Jin, and Jordan 2020),[1] a.k.a., the *last-iterate* convergence, which is a strictly stronger convergence notion and is often more desirable for modern machine learning applications due to the difficulty of averaging agent's actions, typically represented by neural networks. In this paper, we relax the solution concept to the weaker yet arguably fundamental notation of iterative dominance elimination, but obtain last-iterate convergence guarantee in arbitrary games for the first time.

In strategic games, an action a of some agent is *dominated* by another action a' if the agent's payoff of action a is always smaller than her payoff of a', regardless of what actions other agents play. It is widely observed and well accepted that rational human players would avoid playing dominated actions, and such elimination of dominated actions have been observed in human subjective experiments (Fudenberg and Liang 2019). Therefore, an intriguing question is whether algorithms can efficiently replicate such human intelligence. Notably, after eliminating some dominated actions, other actions may then start to become dominated and thus require an additional iteration of dominance elimination. The study of iterative dominance elimination in games dates back to 1950s (Gale 1953; Raiffa and Luce 1957) and has appeared in a variety of applications, including voting (Moulin 1979), auctions (Azrieli and Levin 2011), market design (Abreu and Matsushima 1992), supermodular game (Milgrom and Roberts 1990), oligopolistic competition (Börgers and Janssen 1995) and global games (Carlsson and Van Damme 1993). The method of iterative dominance elimination can be applied to any game for agents to remove "irrational" actions. In situations where the game ends up with only one action profile, the game is said to be *dominance solvable*, and this remaining action profile is its only Nash equilibrium. As we will formally show in this paper, such *iterative dominance elimination* turns out to be a highly non-trivial task and many existing algorithms are provably inefficient.

A celebrated example of iterative dominance elimination is Akerlof's "market for lemons" (Akerlof 1978). Each seller in this market is looking to sell used cars which are equally likely to have quality `H`/high, `M`/medium or `L`/low (a.k.a., "lemons"). Prospective buyers value `H`-cars at \$1000, `M`-cars at \$500 and `L`-cars at \$0, whereas sellers value keeping a `H`-car at \$800, `M`-car at \$400 and `L`-car at \$0 (these values are privy to sellers). Akerlof studies the situation that sellers precisely know their car's type whereas buyers cannot dis-

[*]These authors contributed equally.

[1]In fact, the actions of no-regret agents cannot even stabilize in zero-sum games (Mertikopoulos, Papadimitriou, and Piliouras 2018; Daskalakis and Panageas 2018; Lin, Jin, and Jordan 2020).

tinguish the good cars from lemons. Due to uncertainty in car quality, any buyer will immediately eliminate any price above her average value $\$500 = (1000 + 500 + 0)/3$. After this elimination, the buyer's price becomes lower than H-car's reservation value, and thus drive H-car sellers out of the market. Consequently, the buyer will *gradually learn* that the market has no H-cars under price $\$500$ and thus will *iteratively* eliminate any price above $\$250 = (500 + 0)/2$, which then further drives M-car sellers out of the market. Ultimately, Akerlof observes that this iterative dominance elimination procedure will drive all good cars out of the market, and only lemons are ever traded. In the experiment, we will study a more realistic situation and seek to understand *how fast the market collapse observed by Akerlof may happen when players have such noisy bandit information feedback.*

Another reason that dominance solvability is so fundamental in game theory is due to its connection to the notion of *rationalizability* developed in a series of seminal economic works (Bernheim 1984; Pearce 1984; Milgrom and Roberts 1990; Börgers 1993). In a nutshell, rationalizability is a robust game-theoretic solution concept that strictly generalizes Nash equilibrium (NE) and it turns out that in any 2-player games, all actions that survived iterative dominance elimination are rationalizable (Milgrom and Roberts 1990; Bernheim 1984). We thus believe the goal of learning to eliminate all dominated actions is natural and also essential for any multi-agent systems: not only because without approaching the rationalizability it is impossible to reach the NE or even CE (see Prop. 1), but also rationalizability is often a more realistic outcome to expect in general game-theoretical settings.

An immediate thought one might have is whether any standard no-regret learning algorithm would already suffice to eliminate the (obviously bad) dominated actions. The answer is positive, but with a crucial limitation that they may necessarily take *exponentially* many rounds, as we will prove later. A key insight revealed from our formal results is that the notion of regret in multi-agent settings is not fully aligned with the performance in iterative dominance elimination. First, the history that standard no-regret algorithms exploit could become the inertia that impedes the iterative process of dominance elimination. This claim shall be self-evident in the proof of Theorem 2. Second, the notion of "regret", designed for either stochastic or adversarial settings, fails to encourage the coordination that facilitates the iterative learning process of the learning agents. These observations echoes with the findings of (Viossat and Zapechelnyuk 2013) that the Hannan sets (Hannan 2016) may contain highly non-rationalizable outcomes.

Motivated by the aforementioned fundamentality and intricacies, this paper studies how agents in a multi-agent system can learn to iteratively eliminate all dominated actions under *noisy bandit information* feedback. This is also a natural generalization of the well-known *action elimination* problem (Even-Dar et al. 2006) to multi-agent setups. Our study reveals interesting new challenges of learning in game-theoretical settings that its algorithm design may require different insights from the classical online learning under either adversarial or stochastic environment assumptions.

Contributions. Our contributions are twofold. First, we provide formal barriers about the difficulty of the iterative elimination of all dominated strategies under noisy bandit feedback. To do so, we identify an interesting benchmark class of dominance solvable game instances, coined *diamond-in-the-rough game* (DIR), and show that a broad class of no-regret online learning algorithms, i.e., Dual Averaging (Nesterov 2009; Xiao 2010), has to use exponentially many rounds to eliminate all dominated actions with non-increasing learning rates. Moreover, we prove that the algorithms with the stronger no *swap* regret suffers similar exponential inefficiency. Second, we propose a new variant of the Exp3 algorithm with a carefully designed diminishing history mechanism to overcome the barriers in such learning tasks and prove its efficiency of eliminating all dominated actions within polynomially many rounds in the sense of last-iterate convergence. Our experiments demonstrate the effectiveness of our algorithm not only in the synthetic DIR games but also in other real-world games.

Related Work. Multi-agent learning in games has been of interest since the early days of artificial intelligence and economics (Von Neumann and Morgenstern 2007; Brown 1951; Hart and Mas-Colell 2000). In recent years, there is a growing body of work on decentralized no-regret dynamics and their equilibrium convergence properties in various classes of games including zero-sum game (Daskalakis, Deckelbaum, and Kim 2011; Rakhlin and Sridharan 2013; Syrgkanis et al. 2015; Daskalakis et al. 2017; Daskalakis and Panageas 2018; Mertikopoulos et al. 2018), concave game (Mertikopoulos and Sandholm 2016; Bravo, Leslie, and Mertikopoulos 2018; Mertikopoulos and Zhou 2019), potential games (Cohen, Héliou, and Mertikopoulos 2017b), and auctions (Feng et al. 2021). Different from the goal of these works on convergence to coarse correlated equilibrium or Nash equilibria (for special game classes), we target a different though arguably equally fundamental goal, i.e., learn to *rationalize* by removing dominated actions. Indeed, economists (Viossat 2015) framed this learning goal broadly into the question *whether evolutionary processes lead economic or biological agents to behave as if they were rational.* Our study focuses on the convergence properties of various no-regret learning algorithms specifically in the multi-agent *bandit* learning setting (a.k.a. the "radically uncoupled" setup (Foster and Young 2006)). Interestingly, our proposed mechanism of diminishing history resonates with the well-established studies of both behavioral economy (Fudenberg and Peysakhovich 2016) and political science (Axelrod and Hamilton 1981). It is also seen in one form or another (such as increasing learning rate or recency bias) of many learning algorithms for different purposes (Rakhlin and Sridharan 2013; Syrgkanis et al. 2015; Bubeck, Lee, and Eldan 2017; Agarwal et al. 2017; Lee et al. 2020), some of which even beyond the domain of online learning (Jin et al. 2018; Brown and Sandholm 2019). In the experimental section, we will compare our algorithm with algorithms from these previous works.

Preliminaries

Notations in Multi-Agent Games An N-player game in normal form consists of a (finite) set of agents $\mathcal{N} = \{1, \dots, N\}$, where the n-th agent have a finite set of actions (or pure strategies) $\mathcal{A}_n$. Let $\mathcal{A} \equiv \prod_{n\in\mathcal{N}} \mathcal{A}_n$ denote the action space, $\boldsymbol{a} \equiv (a_1, \dots a_N) \in \mathcal{A}$ denote the action profile, and a_{-n} as the action profile excluding agent-n's action. Without loss of generality, we assume every agent has K actions, i.e., $|\mathcal{A}_n| = K$. Each agent-n has a payoff function $u_n : \mathcal{A} \to [-1, 1]$ that maps the action profile (a_n, a_{-n}) of all agents' actions to the nth agent's reward $u_n(a_n, a_{-n})$. We denote such game instance as $\mathcal{G} \equiv \mathcal{G}(\mathcal{N}, \mathcal{A}, u)$.

The agents may also randomize their actions by playing a *mixed strategy*, i.e. a distributions $x_n \in \Delta_{\mathcal{A}_n}$ over their action sets $\mathcal{A}_n$. We accordingly denote $\mathcal{X}_n \equiv \Delta_{\mathcal{A}_n}$ for the mixed strategy space of agent-n, and $\mathcal{X} \equiv \prod_{n\in\mathcal{N}} \mathcal{X}_n$ as the space of all mixed strategy profiles $x \equiv (x_1, \dots, x_N)$ aggregating over all agents. With slight abuse of notation, we denote the expected payoff of agent-n in the profile x as $u_n(x_n, x_{-n}) = \sum_{a_1\in\mathcal{A}_1} \cdots \sum_{a_N\in\mathcal{A}_N} u_n(a_1, \dots, a_N) \prod_{n\in\mathcal{N}} x_n(a_n)$.

Iterative Dominance Elimination and Dominance Solvable Games. In a game $\mathcal{G}$, if $u_n(x_n, a_{-n}) > u_n(a_n, a_{-n})$, $\forall a_{-n} \in \mathcal{A}_{-n}$, then we say action (i.e., pure strategy) a_n is strictly *dominated* by a *mixed* strategy $x_n \in \Delta_{\mathcal{A}_n}$, or x_n strictly *dominates* a_n. In multi-agent setups, eliminating all dominated actions *iteratively* may require many *iterations* of dominance elimination, as illustrated in the introduction. Moreover, an action dominated by a mixed strategy is not necessarily dominated by any pure strategy. So it is important to consider dominance elimination by mixed strategies. The process of iteratively applying such procedure to remove dominated actions is called *iterated elimination of strictly dominated strategies* (IESDS). This motivates our following natural definition of *elimination length*.

Definition 1. *For any finite game $\mathcal{G}$, we define the **elimination length** L_0 as the minimum number of iterations that IESDS needs to eliminate all dominated actions in $\mathcal{G}$. For any successful execution of IESDS with elimination length L_0, the corresponding **elimination path** is a sequence of **eliminated sets** $(E_l)_{l=1}^{L_0}$ where E_l contains all eliminated actions until iteration $l = 1, \cdots, L_0$.*

By definition, we have $E_1 \subset E_2 \subset \cdots \subset E_{L_0}$ and $|E_{L_0}| < \sum_{i=n}^{N} |\mathcal{A}_n| \le KN$. Let Δ be the smallest utility gap between any dominated action and a correspondingly dominant strategy during IESDS. It has been shown in (Milgrom and Roberts 1990; Bernheim 1984; Pearce 1984) that IESDS can be applied to any finite game to eliminate dominated actions, and in particular, in two-player games, the action profiles that survive the IESDS form the solution concept of *rationalizability*. If IESDS terminates with only a single action profile left in the remaining action space, this action profile must be the unique NE and also the unique CE of the game (Viossat 2008). In this case, game $\mathcal{G}$ is called *mixed-strategy solvable* (Alon, Rudov, and Yariv 2021) — a more general notion than the classic dominance solvable that is defined on dominance elimination by pure strategies.

Multi-Agent Online Learning with Noisy Bandit Feedback At each round $t \in [T]$, each agent-n takes an action $a_n(t)$, which together forms the action profile $\boldsymbol{a}(t)$. Then, agent-n individually observes from the environment a noisy, bandit feedback, $u_n(\boldsymbol{a}(t)) + \xi_{n,t}$, that is, its payoff under the action profile $\boldsymbol{a}(t)$ perturbed by noise $\xi_{n,t}$. With a slight abuse of notation, we denote $u_n(t) = u_n(\boldsymbol{a}(t)) + \xi_{n,t}$ hereinafter. We assume $\{\xi_{n,t}, \mathcal{F}_t\}_0^{+\infty}$ is a Martingale difference sequence (MDS) with finite variance. Specifically, $\{\xi_{n,t}, \mathcal{F}_t\}_0^{+\infty}$ satisfies:

1. Zero-mean: $\mathbb{E}[\xi_{n,t}|\mathcal{F}_{t-1}] = 0$ for all $t = 1, 2, \dots (a.s.)$
2. Finite variance: $\exists\sigma > 0$ s.t. $\mathbb{E}[\|\xi_{n,t}\|^2|\mathcal{F}_{t-1}] \le \sigma^2$ for all $t = 1, 2, \dots (a.s.)$

Learning Goals Given the above problem setting, we focus on designing an online learning algorithm that when all agents use such an algorithm, they will learn to eliminate all dominated actions with high probability in a last-iterate convergence manner. We remark that most of the prior work (Hannan 2016; Freund and Schapire 1999; Daskalakis, Deckelbaum, and Kim 2011; Cherukuri, Gharesifard, and Cortes 2017; Cohen, Héliou, and Mertikopoulos 2017a; Syrgkanis et al. 2015; Mertikopoulos and Zhou 2019; Mazumdar, Ratliff, and Sastry 2020) in multi-agent learning either focus on full information or gradient feedback, or consider the time-average convergence, our learning objective of last-iterate convergence under noisy bandit feedback is a stronger and more stable guarantee.

Barriers of Multi-Agent Learning for Iterative Dominance Elimination

This section presents the intrinsic barriers to multi-agent learning for the iterative elimination of dominated actions. Surprisingly, we show that even for dominance solvable games under full information feedback, standard bandit learning algorithms will necessarily take *exponentially* many rounds to converge to the unique NE. This barrier is already significant even in two-player games. Therefore, for ease of presentation, we shall focus on two-player games with a finite action set $[K]$ and refer to the row player as agent A, column player B, and use indices $i, j \in [1, 2, \cdots, K]$ to denote their pure actions.

"Diamond in the Rough" – A Benchmark Setting for Multi-Agent Learning

We introduce an interesting class of two-player *dominance solvable* games, which is a challenging benchmark setting for multi-agent learning. For reference convenience, we call it "*Diamond in the Rough*", whose meaning should become clear after we define the class of games.

Definition 2 (The Diamond-In-the-Rough (DIR) Games)**.** *A diamond-in-the-rough game (DIR) is a two-player game parameterized by (K, c). Each agent have K actions and utility function*

$$u_1(i,j) = \begin{cases} i/\rho & i \le j+1 \\ -c/\rho & i > j+1 \end{cases}, u_2(i,j) = \begin{cases} j/\rho & j \le i \\ -c/\rho & j > i \end{cases}. \tag{1}$$

where $c > 0$ and $\rho = \max\{K, c\}$ is for normalization purpose. Hence, the payoff matrix of the DIR(K, c) game is given by

$$\frac{1}{\max\{K,c\}}\begin{bmatrix} (1,1) & (1,-c) & (1,-c) & \cdots & (1,-c) \\ (2,1) & (2,2) & (2,-c) & \cdots & (2,-c) \\ (-c,1) & (3,2) & (3,3) & \cdots & (3,-c) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & \\ \vdots & \vdots & \vdots & (K-1,K-1) & (K-1,-c) \\ (-c,1) & (-c,2) & \cdots & (K,K-1) & (K,K) \end{bmatrix}. \tag{2}$$

The DIR game exhibits a "nested" dominance structure, which makes it challenging to play rationally. Specifically, observe that A's action 2 dominates action 1. However, this is not the case for B since if A were to play action 1, B's utility $-c$ of action 2 is significantly worse than utility 1 of action 1. Nevertheless, after A eliminates action 1, B's action 2 starts to dominate B's action 1. This property holds in general for DIR game — B's action $j+1$ dominates her action j *only when* A eliminates his actions $\{1, \cdots, j\}$ and similarly A's action $i+1$ dominates action i *only when* B eliminates her actions $\{1, \cdots, i-1\}$. In the end, the real "diamond" is hidden at the action pair (K, K), which is the best for both agents. However, to find this "diamond", both agents have to sequentially remove all the "rough" actions $1, 2, \cdots, K-1$. This is thus the name "*Diamond in the Rough*".

As we will demonstrate both theoretically and empirically, *the DIR game highlights a fundamental challenge in multi-agent learning* — i.e., whether an agent's action is good or bad depends on what actions her opponents take, and such inter-agent externality makes it challenging to learn the optimal decisions. We end this subsection by summarizing a few useful properties of any DIR(K, c) game.

Proposition 1. *The following properties hold for any DIR(K, c) game:*

1. *The game is dominance solvable by alternatively eliminating the row and column player's action in order $1, 2, ...$, until reaching the last strategy profile (K, K). The elimination length $L_0 = 2K - 2$.*
2. *The strategy profile (K, K) is the unique correlated equilibrium (and thus the unique NE). Moreover, both agents achieve the maximum possible utility at this equilibrium.*

No Swap Regret$\not\Rightarrow$Efficient Dominance Elimination

A celebrated result by (Blum and Mansour 2005) shows that no swap regret algorithms provably converge to ϵ-correlated equilibrium where ϵ is the *average* swap regret. Proposition 1 shows that the DIR game has a unique correlated equilibrium; thus, one might expect that running a no swap regret algorithm would quickly eliminate all the dominated action and converge to the unique equilibrium. More generally, would no swap regret algorithm always eliminate dominated actions in the sense that any dominated actions will be played with a small probability after sufficiently many rounds T? Surprisingly, we show that the answer is NO.

Our following theorem shows that for *any* small ϵ, there exist DIR$(\log(\frac{1}{\epsilon}), c)$ games for some constant c, in which an ϵ-correlated equilibrium may *never* play the unique correlated equilibrium (K, K) and, moreover, will lead to a much smaller utility than the agent's equilibrium utility.

Theorem 1. *For any $\epsilon > 0$ and any DIR(K, c) game satisfying $\log(1/\epsilon) \leq (2K-2)\log(c)$, the game always has an ϵ-correlated-equilibrium which plays the (unique) CE strategy (K, K) with probability 0. Moreover, the welfare of this ϵ-CE is at most $\frac{1+\lceil \log(1/\epsilon)/\log(c)\rceil}{2K}$ fraction of the equilibrium welfare.*

Specifically, by picking $K = \log(1/\epsilon)$ and $\log(c) = 1$, we obtain a DIR$(\log(\frac{1}{\epsilon}), c)$ game which admits an ϵ-EC that will put 0 probability at the unique CE (K, K) and has utility at most half of the players' equilibrium utilities. This may appear quite counter-intuitive at the first glance, since how come an ϵ-EC be so such "far away" from the real and unique CE. This turns out to be due to a subtle difference — that is, the ϵ in "ϵ-CE" is measuring the ϵ difference in *agent utilities*, whereas the "far away" conclusion reflected in Theorem 1 is measuring the true distance between agent's *action probabilities*. Though in the limit as $\epsilon \to 0$, the $\epsilon-$CE will tend to an exact CE, Theorem 1 suggests that agent's strategies and equilibrium utilities can be far from the exact CE even when $\epsilon > 0$ is extremely small compared to the game payoff. Therefore, the fact that an agent does not have much incentive to deviate when at an ϵ-CE does not imply that the action it plays is close to the (exact) CE action profile, neither implies his utility is close to the CE utility. *This insight also explains why classic no regret learning algorithm designed based on maximizing accumulated rewards may perform poorly for the task for iterative dominated action elimination*, which is shown in our next theoretical result.

To concretize the message in Theorem 1, consider a very small DIR with $c = K = 10$. With an $O(\sqrt{T})$ swap regret algorithm, the empirical distribution of the no-regret learning agents is guaranteed to be a $10^{-9}-$CE after $T = 10^{18}$ rounds. However, since $\log(1/\epsilon) = \log(10^9) < (2K-2)\log(c)$, Theorem 1 implies that this $10^{-9}-$CE may still never play the equilibrium strategy (K, K) and its welfare is at most $\frac{1+\lceil \log(1/\epsilon)/\log(c)\rceil}{2K} = \frac{1}{2}$ of the equilibrium welfare. Notably, 10^{18} rounds of sequential interactions would take hundreds of years for a modern CPU to simulate.

Exponential-Time Convergence of Dual Averaging Algorithm

We now show that a broad class of standard no regret algorithms, i.e., Dual Averaging (DA), could fail to eliminate all dominated actions efficiently. The family of DA algorithms includes many celebrated algorithms such as Exponential Weight and lazy gradient descent. In online learning literature, DA is also known as the "lazy" version of online mirror descent (Shalev-Shwartz et al. 2011). The DA algorithm requires the first-order gradient feedback and is parameterized by a learning rate sequence $\{\eta_t\}$ and a "mirror map" Q. The algorithm maintains a score vector $\boldsymbol{y}_t$ to characterize the quality of each action and updates it with the gradient feedback. When choosing the action for the next step, the mirror map Q maps the score vector $\boldsymbol{y}_t = (y_1(t), \cdots, y_K(t)) \in \mathcal{Y}$ to a probability distribution $\boldsymbol{p}_t = (p_1(t), \cdots, p_K(t)) \in \Delta_K$ and randomly samples an action from $\boldsymbol{p}_t$ as the next move. The outline of the DA algorithm is given in Algorithm 1.

Algorithm 1: The Dual Averaging (DA) Algorithm Framework

Input: Mirror map $Q : \mathcal{Y} \to \Delta_K$, learning rate sequence $\{\eta_t > 0\}$
Initialization: $\boldsymbol{y}_1 = (0, \cdots, 0)$,
for $t = 1 \dots T$ **do**
  Compute $\boldsymbol{p}_t = Q(\boldsymbol{y}_t)$.
  Draw an action i_t from the distribution $\boldsymbol{p}_t$.
  Receive the expected payoff $\tilde{\boldsymbol{u}}_t = (\tilde{u}_1(t), \cdots, \tilde{u}_K(t))$ from the first-order oracle.
  Update $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t \tilde{\boldsymbol{u}}_t$.
end for

By convention, the mirror map Q depends on a convex function (or regularizer) h and is defined as

$$Q(y) = \arg\max_{x \in \Delta_K} \{\langle \boldsymbol{y}, \boldsymbol{x} \rangle - h(\boldsymbol{x})\}, \boldsymbol{y} \in \mathcal{Y}. \quad (3)$$

Here are some widely used learning algorithms that happen to be the special cases of DA:

1. when $h(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ is the quadratic function, the mirror map $Q(\boldsymbol{y}) = \arg\min_{\boldsymbol{x} \in \Delta_K} \|\boldsymbol{x} - \boldsymbol{y}\|^2$ takes the form of Euclidean projection and we obtain the lazy gradient descent (LGD) algorithm.
2. when $h(\boldsymbol{x}) = \sum_{i \in [K]} x_i \log x_i$ is the entropic regularizer, the mirror map $Q(\boldsymbol{y}) = \frac{(\exp(y_1), \cdots, \exp(y_K))}{\exp(y_1) + \cdots + \exp(y_1)}$ takes the form of logit choice map and we obtain entropic gradient descent algorithm, which is also known as the Hedge or Exponential Weight (EW).
3. when $h(\boldsymbol{x}) = \frac{1}{p}\|\boldsymbol{x}\|^p$ is the normalized L_p norm and $p \to \infty$, $Q(\boldsymbol{y})$ always returns the pure best strategy to the opponent's average past mixed strategy, and the corresponding algorithm is known as fictitious play (Brown 1951; Viossat and Zapechelnyuk 2013).

We consider the situation where all agents are running DA with typically adopted non-increasing learning rates.[2] That is, at any round t, each agent will do the standard update for any action i using estimated reward $\tilde{u}_i(t)$ with learning rate η_t that is non-increasing in t. Perhaps surprisingly, even with perfect gradient feedback (as oppose to the noisy gradient from bandit feedback)[3], agents following the DA algorithm under certain mild assumptions for the mirror mapping Q will take provably exponential rounds to converge to the unique NE in DIR games.

Theorem 2. *Consider the DIR(K, c) game with any $K \geq 3$ and $c \geq 3K^2$. Suppose two agents both follow the DA algorithm 1 equipped with a mirror map Q defined in Eq* (3) *and a non-negative, bounded,* non-increasing *learning rate sequence $\{\eta_t\}_{t=1}^{\infty}$. Then with a probability at least $1/4$, at least one of the two agents would place zero probability on the (unique) pure NE strategy at any round $t \leq 3^{K-2}$.*

[2]Variants of EW (a.k.a., multiplicative weight updates) have been proved to converge to equilibria in other games such as potential games (Cohen, Héliou, and Mertikopoulos 2017b) and concave games (Bravo, Leslie, and Mertikopoulos 2018).

[3]The perfect gradient of each agent n in the game is a vector with each entry, $\tilde{u}_{a_n}(t) = \sum_{a_{-n}} p_{a_{-n}}(t) u_n(a_n, a_{-n})$, equivalent to the expected payoff of each action a_n given the opponents' strategy profile, since the loss is a linear function of the payoff. The noisy gradient estimated from bandit feedback can be found in Algorithm 2.

Algorithm 2: `Exp3` with `D`iminishing `H`istory (`Exp3-DH`)

Input: Number of actions K, parameter ϵ_t, β .
Initialization: $y_i(0) = 0, \forall i \in [K]$.
for $t = 0, 1, \dots, T$ **do**
  Set

$$p_i(t) = (1 - \epsilon_t)\frac{\exp(y_i(t))}{\sum_{j \in [K]} \exp(y_j(t))} + \frac{\epsilon_t}{K} \quad \forall i \in [K].$$

  Draw action i_t with prob. $(p_1(t), p_2(t), \dots p_K(t))$.
  Observe realized payoff $u_{i_t}(t)$.
  Compute *unbiased* payoff estimator $\tilde{u}_i(t) = \frac{u_i(t)}{p_i(t)} \cdot \mathbb{I}(i = i_t), \forall i \in [K]$.
  Update

$$y_i(t+1) = (\frac{t}{t+1})^{\beta} \cdot y_i(t) + \tilde{u}_i(t), \quad \forall i \in [K].$$

end for

We emphasize that our negative result holds for any generic mirror map Q and is thus applicable to all widely used special cases (e.g., LGD, EW, etc.). Thus, Theorem 2 proves the inefficiency of any DA algorithm with a non-increasing learning rate sequence to efficiently eliminate all dominated strategies in DIR. We remark that the requirement $c \geq 3K^2$ is only necessary to the proof technique and does not imply the DIR game is easy to solve when $c < 3K^2$. As we demonstrate in numerical experiments, the choice of $c = O(K)$ already results in an extremely slow convergence rate empirically for Exp3 algorithm and its variants. We also emphasize that Theorem 2 does not contradict previous positive convergence results in (Cohen, Héliou, and Mertikopoulos 2017a; Mertikopoulos and Zhou 2019; Laraki and Mertikopoulos 2013). Please refer to the remarks on Theorem 2 in our full paper (Wu, Xu, and Yao 2021) for detailed discussions.

`Exp3-DH` and its Efficiency for Iterative Dominance Elimination

Motivated by the barriers in Section , we now introduce a novel algorithm *Exp3 with Diminishing History* (`Exp3-DH`) described in Algorithm 2, which turns out to provably guarantee efficient iterative elimination of all dominated actions, with high probability.

`Exp3-DH` is an EW-style algorithm but with important characteristics as follows: 1.`Exp3-DH` uses unbiased payoff estimator $\tilde{u}_{i_t}(t) = \frac{u_{i_t}(t)}{p_{i_t}(t)}$; 2. during reward updates at Step 8, `Exp3-DH` will always discount *each action*'s previous reward $y_i(t)$ by a factor $(\frac{t-1}{t})^{\beta}$ for some carefully chosen parameter β, *regardless of whether this action is taken at this round or not*. Hence, rate of historical rewards will gradually diminish, which thus leads to the name of our algorithm.

Effective Learning Rates. Note that the update in Step 8 of Algorithm 2 only captures the recursive relation between $y_i(t+1)$ and $y_i(t)$. From this recursion, we can easily derive how $y_i(t)$ depends on all previous payoff estimation $\tilde{u}_i(\tau)$ for $\tau = 0, 1, \cdots, t$, which is the follows,

$$y_i(t+1) = \sum_{\tau=0}^{t} (\frac{\tau}{t})^{\beta} \tilde{u}_i(\tau). \tag{4}$$

For notational convenience, we call $\gamma_\tau^{(t)} = (\tau/t)^\beta$ the *effective* learning rate. Notably, the learning rate $\gamma_\tau^{(t)}$ for any fixed past payoff estimation $\tilde{u}_i(\tau)$ *dynamically decreases* as the round t becomes large. This means that the weight of the historical estimation $\tilde{u}_i(\tau)$ will become smaller and smaller as time goes. In other words, the algorithm exhibits *recency bias* and gradually "forgets" histories and always relies more on recent payoff estimations.

We end this section by comparing `Exp3-DH` with previous Exp3-style algorithms. In standard Exp3 algorithm (Auer et al. 2002), the *effective* learning rate is exactly its learning rate γ_t, which is set to a constant or a decreasing sequence to guarantee a sub-linear regret. Some other variants of Exp3 use non-increasing effective learning rate γ_t (Neu 2015; Mertikopoulos and Zhou 2019; Cohen, Héliou, and Mertikopoulos 2017b; Bravo, Leslie, and Mertikopoulos 2018). However, `Exp3-DH` differs from these algorithms in at least two key aspects: (1) its effective learning rate $\gamma_\tau^{(t)}$ is increasing in τ, i.e., biased towards recent reward estimations; (2) $\gamma_\tau^{(t)}$ will be re-scaled every time t increases, i.e., a new observation comes. The first deviation is justified by our Theorem 2 since DA with any decreasing learning rate necessarily suffer exponential convergence. We note that increasing learning rate has been recently studied for single-agent bandit problems (Bubeck, Lee, and Eldan 2017; Lee et al. 2020; Agarwal et al. 2017) and for faster convergence to coarse correlated equilibrium in games (Syrgkanis et al. 2015). Unfortunately, our experimental results show that these algorithms fail to efficiently eliminate all dominated actions, which illustrates that careful design of learning rate is necessary for efficient dominance elimination.

Efficient Convergence of `Exp3-DH`

We now present the theoretical guarantee for `Exp3-DH`.

Definition 3 (*ε-Essential Elimination*)**.** *We say that an action $i \in \mathcal{A}_n$ is* ε-essentially eliminated *at time step T if agent-n's probability of playing i satisfies $p_i(T) \leq \frac{\varepsilon}{4KN}$.*

Note that if all the actions in E_{L_0} are *essentially eliminated* at time step T, the mixed strategy $x(T)$ given by the probability distribution $p(T)$ of all agents satisfies $\|x(T) - x^*\|_1 \leq \frac{\varepsilon}{4KN} \cdot 2(KN - N) < \frac{\varepsilon}{2}$. Therefore, *$\varepsilon$-Essential Elimination* implies last-iterate convergence. We are now ready to give the convergence guarantee for `Exp3-DH` in Theorem 3:

Theorem 3 (Informal)**.** *If all agents run `Exp3-DH` on game $\mathcal{G}(\mathcal{N}, \mathcal{A}, u)$ with parameters $\beta > 0$ and $\epsilon_t = t^{-\frac{1}{3}}$, then after $\tilde{O}(\max\{N^3, K^{1.5}\} \max\{\varepsilon^{-3}, \Delta^{-3}\}(1+\sigma^3)\beta^{1.5} \log^{1.5} \frac{1}{\delta})$ number of rounds, all the dominated actions in $\mathcal{G}$ will be* ε-essentially eliminated *with probability at least $1-\delta$.*

Theorem 3 gives an explicit upper bound for the number of rounds needed for iterative dominance elimination and also captures the convergence rate of ϵ as a function of time horizon T, given $b = 1/3$. Specifically, $\epsilon = O(1/\sqrt[3]{T})$ with the caveat that omitted constants in the big O depends on the game parameters N, K, Δ, δ. If we choose $b = 1/2$, the convergence rate will be $\epsilon = O(1/\sqrt{T})$, however its dependence on the game parameters will be worse. The formal theorem can be found in our full arXiv version (Wu, Xu, and Yao 2021).

Empirical Evaluations

Baselines We compare `Exp3-DH` with a carefully selected family of online learning algorithms: (a) the classical Exp3 algorithm; (b) Exp3.P which has no regret with high probability (Auer et al. 2002); (c) Exp3-RVU which uses recency bias to obtain faster convergence to coarse correlated equilibria in games (Syrgkanis et al. 2015); (d) no-swap regret variant of Exp3.P, coined Exp3.P-swap, which guarantees convergence to $\epsilon-$CE (Blum and Mansour 2005); (e) the online mirror descent algorithm with log barrier regularizer, OMD-LB (Foster et al. 2016), which also uses increasing learning rate schedule (Lee et al. 2020; Agarwal et al. 2017; Bubeck, Lee, and Eldan 2017). We let all learning agents follow the same type of learning algorithm, i.e., self-play, in games described below and compare their convergence trend.

Metrics To measure the progress of iterative dominance elimination, we define the notion of *elimination distance* (ED) for any action i, $\Lambda(i)$, as the number of elimination iterations needed before this action start to be eliminated. Formally, $\Lambda(i) \equiv \arg\max_{0 \leq l \leq L_0} \{i \notin E_l\}$, where $E_0 = \varnothing$. The elimination distance of any undominated action is L_0, the elimination length (see Definition 1). We let $\sum_{i \in \mathcal{A}_n} x_n(i) \frac{\Lambda(i)}{L_0}$ be the normalized ED of mixed strategy x_n of any agent n. We then introduce the *Progress of Elimination* (PoE) metric, as the normalized elimination distance aggregated over all agents in the game at round t,

$$\text{PoE}(t) = \frac{1}{N} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{A}_n} p_{n,i}(t) \cdot \frac{\Lambda(i)}{L_0} \in [0, 1]$$

Therefore, the larger PoE is, the more actions the learning agents have eliminated. When PoE reaches 1, the learning agents have removed all dominated actions and converged to the desirable set of rationalizable actions.

DIR game To illustrate our theoretical results, we conduct a set of experiments in the DIR game, where all agents following the same type of learning algorithm. For DIR game, we use $b = 0.2, \beta = 2K \approx L_0$ as the parameter of `Exp3-DH`. In Figure 1, we can observe that `Exp3-DH` exhibits a superior performance in both cases, and enjoys a greater advantage in a larger game instance, where the other baselines even struggle to eliminate the first few dominated actions. In addition, OMD-LB with increasing learning rate

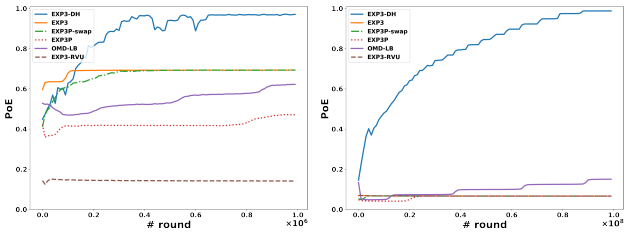

Figure 1: Progress of Elimination (PoE) in a smaller **DIR**$(10, 20)$ **game (left)** over $T = 10^6$ rounds and a larger **DIR**$(20, 40)$ **game (right)** over $T = 10^8$ rounds. In both games, i.i.d. Gaussian noise with std. 0.1 is added onto agents' payoffs. The performance of `Exp3-DH` is represented by blue solid line while five baseline algorithms are represented by other notations shown in the legend.

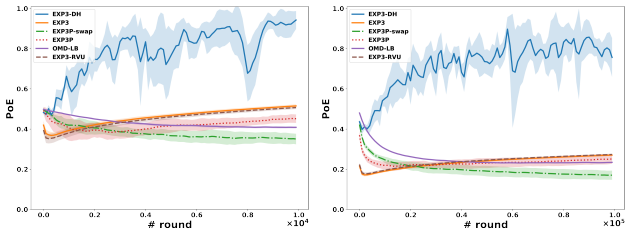

Figure 2: Progress of Elimination (PoE) in **The Market for "Lemons"** with 50 sellers (Left) or 200 sellers (Right). In this game, i.i.d. Gaussian noise with std. 0.1 is added onto agents' payoffs. The lightly shaded region displays the error bar of each convergence trend (by one standard deviation over 5 runs).

also displays relatively good performance especially in the larger and harder instance, compare to other learning algorithms with non-increasing learning rate. But in the harder instance of DIR game with just 20 actions on the right, all the baseline algorithms can only do elimination for 5 or 6 iterations out of $2K-2 = 38$ iterations. We also demonstrate the learning dynamics in these two instances through the animation at https://github.com/lab-sigma/learning-to-rationalize.

The Market for "Lemons" We also examine the algorithm performance on the famous example of the adverse selection problem in (Akerlof 1978). Formally, consider a market of used cars with a buyer and N sellers. Each seller i has a car of quality q_i, and two actions $a_i \in \{1, 0\}$, respectively, to list or not to list his car. Without loss of generality, let $q_N > q_{N-1} > \cdots > q_1$. The buyer has her action $p \in \mathcal{P}$ from a set of prices to buy a car from sellers. Suppose the buyer and sellers move simultaneously. As assumed by Akerlof, the buyer has no information about each seller's car quality before posting the price. The seller also decides whether or not to list his car (i.e., $a_i = 1$ or 0) without knowing the price. In our experiments, we assume seller i has a reservation value $\tilde{q}_i = q_i + \epsilon$ which is a noisy perception of his car quality with zero-mean noise ϵ. For those who did choose to list their cars, if the buyer's price are below their reservation value, they would refuse to sell, but still suffers a small and fixed opportunity cost c_1. We denote $b_i = \mathbf{1}[\tilde{q}_i \le p]$ as whether seller i's car gets sold. In contrast, the buyer is uninformed of the car quality he could buy, though the trade would generate welfare so that her revenue is a multiplier c_2 of the average quality $\bar{q} = \frac{\sum_{i\in[N]} b_i \cdot q_i}{\sum_{i\in[N]} b_i}$. Hence, buyer receives utility $u_0 = c_2 \cdot \bar{q} - p$, and each seller i receives utility $u_i = a_i(b_i(p - q_i) - c_1)$, where $c_1 > 0, c_2 > 1$ are game parameters.

In our full paper, we formally show that this game has elimination length at least $2N - 1$ for small c_1 and its Nash equilibria are that the buyer sets a price no higher than q_1, and all sellers refuse to list. This equilibirum outcome is certainly not a surprise and was observed by Akerlof as a *market collapse*, due to asymmetric information among sellers and buyers which gradually drives the high quality car sellers out of the market in order. The additional insight from this game is that this market collapse may follow a long procedure of iterative dominance elimination. We remark that the opportunity cost $c_1 > 0$ is only assumed to capture how much sellers prefer not selling if the price just matches their values. Our results hold for $c_1 = 0$ as well, but will need to work with weakly dominance elimination with a tie breaking in favor of not listing.

Our numerical experiments aim at testing how fast the market will collapse, if every buyer i have noisy and bandit perception $\tilde{q}_i$ of their car's true quality. In our experiment instances, we set $c_1 = 3, c_2 = 1.5$, i.i.d. noise $\epsilon \sim \mathcal{N}(0, 5)$. Respectively, we choose $N = K - 1 = 50$ or 200 and let each $q_i = N/2 + i$, $\mathcal{P} = \{N/2, \cdots, 3N/2\}$. We let the buyer and sellers apply no-regret learning algorithm to learn the equilibrium price and listing decisions from only the noisy bandit feedback in a repeated game. For `Exp3-DH`, we set $b = 0.5, \beta \approx L_0$, since N is of the same order with K in this game. In Figure 2, as predicted by our theoretical results, with all learning agents running `Exp3-DH` algorithm, the convergence can be polynomial, whereas the convergence from existing no-regret learning algorithm are extremely slow. Interestingly, algorithms such as the EXP3P-swap and OMG-LB even tends to move away from the market collapse outcome.

Conclusion

Our study formalizes the price of "over-hedging" of standard no-regret learning algorithms in the process of iterative dominance elimination. Such price is especially expensive in games whose equilibria are hidden in the "rough". In order to overcome this pitfall, we design a diminishing-history mechanism that deliberately balance the exploitation of the existing knowledge and indifference to history. While it remains unclear whether online mirror descent (OMD), another family of no regret learning algorithm, suffers the similar learning barrier, this work embarks on the fundamental question about the existence of an intrinsic gap between the no-regret learning objective and the convergence to equilibrium in multi-agent learning.

References

Abreu, D.; and Matsushima, H. 1992. Virtual implementation in iteratively undominated strategies: complete information. *Econometrica: Journal of the Econometric Society*, 993–1008.

Agarwal, A.; Luo, H.; Neyshabur, B.; and Schapire, R. E. 2017. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 12–38. PMLR.

Akerlof, G. A. 1978. The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, 235–251. Elsevier.

Alon, N.; Rudov, K.; and Yariv, L. 2021. Dominance Solvability in Random Games. arXiv:2105.10743.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77.

Axelrod, R.; and Hamilton, W. D. 1981. The evolution of cooperation. *science*, 211(4489): 1390–1396.

Azrieli, Y.; and Levin, D. 2011. Dominance-solvable common-value large auctions. *Games and Economic Behavior*, 73(2): 301–309.

Bernheim, B. D. 1984. Rationalizable strategic behavior. *Econometrica: Journal of the Econometric Society*, 1007–1028.

Blum, A.; and Mansour, Y. 2005. From external to internal regret. In *International Conference on Computational Learning Theory*, 621–636. Springer.

Börgers, T. 1993. Pure strategy dominance. *Econometrica: Journal of the Econometric Society*, 423–430.

Börgers, T.; and Janssen, M. C. 1995. On the dominance solvability of large Cournot games. *Games and Economic Behavior*, 8(2): 297–321.

Bravo, M.; Leslie, D.; and Mertikopoulos, P. 2018. Bandit learning in concave N-person games. In *Advances in Neural Information Processing Systems*, 5661–5671.

Brown, G. W. 1951. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1): 374–376.

Brown, N.; and Sandholm, T. 2019. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1829–1836.

Bubeck, S.; Lee, Y. T.; and Eldan, R. 2017. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 72–85.

Carlsson, H.; and Van Damme, E. 1993. Global games and equilibrium selection. *Econometrica: Journal of the Econometric Society*, 989–1018.

Cherukuri, A.; Gharesifard, B.; and Cortes, J. 2017. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1): 486–511.

Cohen, J.; Héliou, A.; and Mertikopoulos, P. 2017a. Hedging under uncertainty: regret minimization meets exponentially fast convergence. In *International Symposium on Algorithmic Game Theory*, 252–263. Springer.

Cohen, J.; Héliou, A.; and Mertikopoulos, P. 2017b. Learning with bandit feedback in potential games. In *Proceedings of the 31th International Conference on Neural Information Processing Systems*.

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

Daskalakis, C.; Ilyas, A.; Syrgkanis, V.; and Zeng, H. 2017. Training gans with optimism. *arXiv preprint arXiv:1711.00141*.

Daskalakis, C.; and Panageas, I. 2018. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, 9236–9246.

Even-Dar, E.; Mannor, S.; Mansour, Y.; and Mahadevan, S. 2006. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of machine learning research*, 7(6).

Feng, Z.; Guruganesh, G.; Liaw, C.; Mehta, A.; and Sethi, A. 2021. Convergence Analysis of No-Regret Bidding Algorithms in Repeated Auctions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5399–5406.

Foster, D.; and Young, H. P. 2006. Regret testing: Learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1(3): 341–367.

Foster, D. J.; Li, Z.; Lykouris, T.; Sridharan, K.; and Tardos, E. 2016. Learning in games: Robustness of fast convergence. *arXiv preprint arXiv:1606.06244*.

Foster, D. P.; and Vohra, R. 1999. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2): 7–35.

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

Fudenberg, D.; and Liang, A. 2019. Predicting and understanding initial play. *American Economic Review*, 109(12): 4112–41.

Fudenberg, D.; and Peysakhovich, A. 2016. Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem. *ACM Transactions on Economics and Computation (TEAC)*, 4(4): 1–18.

Gale, D. 1953. A theory of n-person games with perfect information. *Proceedings of the National Academy of Sciences of the United States of America*, 39(6): 496.

Hannan, J. 2016. 4. APPROXIMATION TO RAYES RISK IN REPEATED PLAY. In *Contributions to the Theory of Games (AM-39), Volume III*, 97–140. Princeton University Press.

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

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 4863–4873.

Laraki, R.; and Mertikopoulos, P. 2013. Higher order game dynamics. *Journal of Economic Theory*, 148(6): 2666–2695.

Lee, C.-W.; Luo, H.; Wei, C.-Y.; and Zhang, M. 2020. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs. *Advances in Neural Information Processing Systems*, 33.

Lin, T.; Jin, C.; and Jordan, M. 2020. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, 6083–6093. PMLR.

Mazumdar, E.; Ratliff, L. J.; and Sastry, S. S. 2020. On Gradient-Based Learning in Continuous Games. *SIAM Journal on Mathematics of Data Science*, 2(1): 103–131.

Mertikopoulos, P.; Lecouat, B.; Zenati, H.; Foo, C.-S.; Chandrasekhar, V.; and Piliouras, G. 2018. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*.

Mertikopoulos, P.; Papadimitriou, C.; and Piliouras, G. 2018. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2703–2717. SIAM.

Mertikopoulos, P.; and Sandholm, W. H. 2016. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4): 1297–1324.

Mertikopoulos, P.; and Zhou, Z. 2019. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1): 465–507.

Milgrom, P.; and Roberts, J. 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica: Journal of the Econometric Society*, 1255–1277.

Moulin, H. 1979. Dominance solvable voting schemes. *Econometrica: Journal of the Econometric Society*, 1337–1351.

Nesterov, Y. 2009. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1): 221–259.

Neu, G. 2015. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances on Neural Information Processing Systems 28 (NIPS 2015)*, 3150–3158.

Pearce, D. G. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica: Journal of the Econometric Society*, 1029–1050.

Raiffa, H.; and Luce, R. D. 1957. *Games and Decisions: Introduction and Critical Survey*. John Wiley, New York.

Rakhlin, A.; and Sridharan, K. 2013. Optimization, learning, and games with predictable sequences. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, 3066–3074.

Shalev-Shwartz, S.; et al. 2011. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2): 107–194.

Syrgkanis, V.; Agarwal, A.; Luo, H.; and Schapire, R. E. 2015. Fast Convergence of Regularized Learning in Games. *Advances in Neural Information Processing Systems*, 28: 2989–2997.

Viossat, Y. 2008. Is having a unique equilibrium robust? *Journal of Mathematical Economics*, 44(11): 1152–1160.

Viossat, Y. 2015. Evolutionary dynamics and dominated strategies. *Economic Theory Bulletin*, 3(1): 91–113.

Viossat, Y.; and Zapechelnyuk, A. 2013. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2): 825–842.

Von Neumann, J.; and Morgenstern, O. 2007. *Theory of games and economic behavior (commemorative edition)*. Princeton university press.

Wu, J.; Xu, H.; and Yao, F. 2021. Multi-Agent Learning for Iterative Dominance Elimination: Formal Barriers and New Algorithms. *arXiv preprint arXiv:2111.05486*.

Xiao, L. 2010. Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization. *Journal of Machine Learning Research*, 11(88): 2543–2596.