Competing Against Equilibria in Zero-Sum Games with Evolving Payoffs

Adrian Rivera Cardoso [1] **Jacob Abernethy** [2] **He Wang** [1] **Huan Xu** [1]

Abstract

We study the problem of repeated play in a zero-sum game in which the payoff matrix may change, in a possibly adversarial fashion, on each round; we call these Online Matrix Games. Finding the Nash Equilibrium (NE) of a two player zero-sum game is core to many problems in statistics, optimization, and economics, and for a fixed game matrix this can be easily reduced to solving a linear program. But when the payoff matrix evolves over time our goal is to find a sequential algorithm that can *compete with*, in a certain sense, the NE of the long-term-averaged payoff matrix. We design an algorithm with small NE regret–that is, we ensure that the long-term payoff of both players is close to minimax optimum in hindsight. Our algorithm achieves near-optimal dependence with respect to the number of rounds and depends poly-logarithmically on the number of available actions of the players. Additionally, we show that the naive reduction, where each player simply minimizes its own regret, fails to achieve the stated objective regardless of which algorithm is used. Lastly, we consider the so-called *bandit setting*, where the feedback is significantly limited, and we provide an algorithm with small NE regret using one-point estimates of each payoff matrix.

1. Introduction

We consider a problem in which two players interact in a zero-sum game repeatedly. The payoff matrix of the game is unknown to the players *a priori*, and may change arbitrarily on each round. Our objective is to find competitive strategies that can achieve the Nash equilibrium of the game with the average payoffs in the long term. This problem is a significant extension of the classical learning setting in zero-sum games, where the underlying payoff matrix is often assumed to be fixed or i.i.d. In contrast, we allow the payoff matrix to evolve arbitrarily in each round, and can even be selected in a possibly adversarial fashion.

Zero-sum games (Von Neumann, 1928; Morgenstern & Von Neumann, 1953) are ubiquitous in economics and central to understanding Linear Programming duality (Hazan, 2016; Adler, 2013), convex optimization (Abernethy & Wang, 2017; Abernethy et al., 2018), robust optimization (Ben-Tal et al., 2009), and Differential Privacy (Dwork et al., 2014). The task of finding the Nash equilibrium of a zero-sum game is also connected to several machine learning problems such as: Markov Games (Littman, 1994), Boosting (Freund & Schapire, 1996), Multiarmed Bandits with Knapsacks (Badanidiyuru et al., 2013; Immorlica et al., 2018) and dynamic pricing problems (Ferreira et al., 2018).

We formally define the problem setting in Section 1.1. We then highlight the main contributions of this paper in Section 1.2 and discuss related works in Section 1.3.

1.1. Problem Formulation: Online Matrix Games

We start by reviewing the definition of classical two-player zero-sum games. Suppose player 1 has d_1 possible actions and player 2 has d_2 possible actions. The payoffs for both players are determined by a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, with $A_{i,j}$ corresponding to the loss of player 1 and the reward of player 2 when they choose to play actions $(i,j) \in [d_1] \times [d_2]$.[1] We allow the players to use *mixed strategies* – each mixed strategy is represented by a probability distribution over their actions. More specifically, when Player 1 uses a mixed strategy $x \in \Delta_{d_1}$ and Player 2 uses a mixed strategy $y \in \Delta_{d_2}$, the expected payoff is $x^\top Ay$.[2] Throughout the paper, we refer to the static zero-sum game as a *matrix game* (MG), because the players' payoffs are a bilinear function encoded by the matrix A. A Nash equilibrium of this game is defined as any pair of (possibly) mixed strategies (x^*, y^*) such that

$$(x^*)^\top Ay \leq (x^*)^\top Ay^* \leq x^\top Ay^*$$

for any $x \in \Delta_{d_1}, y \in \Delta_{d_2}$. It is well known that every MG has at least one Nash equilibrium (Morgenstern & Von Neu-

[1] Department of Industrial and Systems Engineering, Georgia Institute of Technology, GA, USA [2] Department of Computer Science, Georgia Institute of Technology, GA, USA. Correspondence to: Adrian Rivera Cardoso <adrian.riv(at)gatech.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] Throughout, $[n] \triangleq \{1, ..., n\}$ for any positive integer n.

[2] Here, Δ_d represents the unit simplex in dimension d: $\Delta_d \triangleq \{v \in \mathbb{R}^d : \|v\|_1 = 1, v \geq 0\}$.

mann, 1953). The problem of finding an equilibrium for a MG can be reduced to solving linear programming problems. In fact, Adler (2013) showed that the opposite is also true, every linear programming problem can be solved by finding an equilibrium to a corresponding MG.

Now, we define a problem that generalizes the matrix games into an online setting, which we call the Online Matrix Games (OMG) problem. Suppose two players interact in a repeated zero-sum matrix game through T rounds. In every round $t \in [T]$, they must each choose a (possibly) mixed strategy from the given action sets $x_t \in \Delta_{d_1}, y_t \in \Delta_{d_2}$. However, we assume that the payoff matrix in OMG can evolve in each round, and the players have no knowledge of the payoff quantities in that round before they commit to an action. Let $\{A_t\}_{t=1}^T$ be an arbitrary sequence of matrices, where each $A_t \in [-1,1]^{d_1 \times d_2}$ for all $t = 1, .., T$. For each round t, the players choose their mixed strategies $x_t \in \Delta_{d_1}, y_t \in \Delta_{d_2}$ before the matrix A_t is revealed. Then, player 1 (resp. player 2) receives a loss (resp. gain) given by the payoff quantity $x_t^\top A_t y_t$. Note that the payoff matrix A_t is allowed to change arbitrarily from round to round and may even depend on the past actions of both players. The *joint goal* for both players is to find strategies that ensure their average payoffs in T rounds is close to the Nash Equilibrium under the average payoff matrix $\frac{1}{T}\sum_{t=1}^T A_t$ in hindsight.

More precisely, let us call the quantity

$$\left| \sum_{t=1}^T x_t^\top A_t y_t - \min_{x \in X} \max_{y \in Y} \sum_{t=1}^T x^\top A_t y \right| \tag{1}$$

the *Nash Equilibrium (NE) regret*. This is a natural extension of the regret concept in typical online learning or multi-armed bandit problems, which involve only a single decision maker. The primary objective of the OMG problem is to find online strategies for both players so that, as $T \to \infty$, the average NE regret (1) per round tends to 0 (i.e., the NE regret is $o(T)$).

We make some remarks about the choice of benchmark and the fact that the players must update jointly despite the fact that they are playing a zero-sum game. In the following examples, the comparator term $\min_{x \in X} \max_{y \in Y} \sum_{t=1}^T x^\top A_t y$ arises naturally and there is one decision maker which chooses the actions of both players.

1. Online Linear Programming (Agrawal et al., 2014): the decision maker solves an LP where data arrives sequentially. This problem has real-world applications in ad-auctions. Using Lagrangian duality, we can reduce this problem to an online zero-sum game (our setting), where player 1 chooses primal variables and player 2 chooses dual variables. Our benchmark corresponds to the optimal solution of the offline LP.
2. Adversarial Bandits with Knapsacks (Immorlica et al., 2018): this problem extends the classical Multi Armed Bandit by adding a 'knapsack' constraint. Again, using a Lagrangian relaxation on the knapsack constraint, this problem can be linked to the online min-max games that we study (see Sec. 3.2 of (Immorlica et al., 2018)).
3. Generative Adversarial Networks (Goodfellow et al., 2014): GANs can also be viewed as a zero-sum game, where the decision maker trains the generator and discriminator to find a Nash equilibrium. Although our model cannot directly be used for GANs because they are nonconvex, it is another example where both players may desire to update jointly.

In the paper, we consider the OMG problem in two distinct information feedback settings. In the *full information* setting (Section 4), both players are able to observe the full matrix A_t at the end of round t. In the *bandit setting* (Section 5), players can only observe the entry of A_t indexed by (i_t, j_t) at the end of round t, where i_t and j_t are the actions sampled from the probability distributions associated with their mixed strategies (x_t, y_t).

1.2. Main Contributions

In addition to introducing a novel problem setting, the main contributions of the present work are as follows.

- First, we show that a natural "naïve" approach, where each player simply aims to minimize their individual regret, will fail to produce a sublinear NE regret algorithm, in the sense of (1), regardless of the players' no-regret strategies (Theorem 1).
- Second, in the full information setting, we provide an algorithm for the OMG problem that achieves a NE regret of $O(\max\{\ln(d_1), \ln(d_2)\} \ln(T)\sqrt{T})$ (Theorem 3). Note that the regret depends logarithmically on the number of actions, allowing us to handle scenarios where the players have exponentially many actions available.
- Third, we propose an algorithm for the bandit setting that achieves an NE regret of order $O((\max\{d_1, d_2\})^{5/3} T^{5/6})$ (Theorem 5).

1.3. Related Work

The reader familiar with Online Convex Optimization (OCO) may find it closely related to the OMG problem. In the OCO setting, a player is given a convex, closed, and bounded action set X, and must repeatedly choose an action $x_t \in X$ before the convex function $f_t(x) : X \to \mathbb{R}$ is revealed. The player's goal is to obtain sublinear *individual regret* defined as $\sum_{t=1}^T f_t(x_t) - \min_{x \in X} \sum_{t=1}^T f_t(x)$. This problem is well studied and several algorithms such as Online Gradient Descent (Zinkevich, 2003), Regularized Follow the Leader (Shalev-Shwartz & Singer, 2007; Abernethy et al., 2009) and Perturbed Follow the Leader (Kalai & Vempala, 2002) achieve optimal individual regret bounds that scale as $O(\sqrt{T})$. The most natural (although incorrect) approach to attack the OMG problem is to equip each of the players with a sublinear individual regret algorithm. However, we will show in Section 3 that if both players use an algorithm that guarantees sublinear individual regret, then it is impossible to achieve sublinear NE regret when the payoff matrices are chosen adversarially. In other words, the algorithms for the OCO setting cannot be directly applied to the OMG problem considered in this paper.

We now discuss some related works that focus on learning in games. Singh et al. (2000) study a two player, two-action general sum static game. They show that if both players use Infinitesimal Gradient Ascent, either the strategy pair will converge to a Nash Equilibrium (NE), or even if they do not, then the average payoffs are close to that of the NE. A result of similar flavor was derived in Cesa-Bianchi et al. (2007) for any zero-sum convex-concave game. Given a payoff function $\mathcal{L}(x, y)$, they show that if both players minimize their individual-regrets, then the average of actions $(\bar{x}, \bar{y})$ will satisfy $|\mathcal{L}(\bar{x}, \bar{y}) - \mathcal{L}(x^*, y^*)| \to 0$ as $T \to \infty$, where (x^*, y^*) is a NE. Bowling & Veloso (2001) improve upon the result of Singh et al. (2000) by proposing an algorithm called WoLF (Win or Learn Fast), which is a modification of gradient ascent; they show that the iterates of their algorithm indeed converge to a NE. Conitzer & Sandholm (2007) further improve the results in Singh et al. (2000) and Bowling (2005) by developing an algorithm called GIGA-WoLF for multi-player nonzero sum static games. Their algorithm learns to play optimally against stationary opponents; when used in self-play, the actions chosen by the algorithm converge to a NE. More recently, Balduzzi et al. (2018) studied general multi-player static games and show that by decomposing and classifying the second order dynamics of these games, one can prevent cycling behavior to find NE. We note that unlike our paper, all of the papers above consider repeated games with a static payoff matrix, whereas we allow the payoff matrix to change arbitrarily. An exception is the work by Ho-Nguyen & Kılınç-Karzan (2016), who consider the same setting as our OMG problem; however their paper only shows that the sum of the individual regrets of both players is sublinear and does not study convergence to NE.

Related to the OMG problem with bandit feedback is the seminal work of Flaxman et al. (2005). They provide the first sublinear regret bound for Online Convex Optimization with bandit feedback, using a one-point estimate of the gradient. The one-point gradient estimate used in Flaxman et al. (2005) is similar to those independently proposed in Granichin (1989) and in Spall (1997). The regret bound provided in Flaxman et al. (2005) is $O(T^{3/4})$, which is suboptimal. In Abernethy et al. (2009), the authors give the first $O(\sqrt{T})$ bound for the special case when the functions are linear. More recently, Hazan & Li (2016) and Bubeck et al. (2016) designed the first efficient algorithms with $\tilde{O}(poly(d)\sqrt{T})$ regret for the general online convex optimization case; unfortunately, the dependence on the dimension d in the regret rate is a very large polynomial. Our one-point matrix estimate is most closely related to the random estimator in Auer et al. (1995) for linear functions. It is possible to use the more sophisticated techniques from Abernethy et al. (2009); Hazan & Li (2016); Bubeck et al. (2016) to improve our NE regret bound in section 5; however, the result does not seem to be immediate and we leave this as future work.

2. Preliminaries

In this section we introduce notation and definitions that will be used throughout the paper.

2.1. Notation

By default, all vectors are column vectors. A vector with entries $x_1, ..., x_d$ is written as $x = [x_1; ...; x_d] = [x_1, ..., x_d]^\top$, where $\top$ denotes the transpose. For a matrix A, let A_{ij} be the entry in the i-th row and j-th column.

2.2. Convex Functions

For any $H > 0$ we say that a function $f : X \to \mathbb{R}$ is H-strongly convex with respect to a norm $\|\cdot\|$, if for any $x_1, x_2 \in X$, it holds that

$$f(x_1) \geq f(x_2) + \nabla f(x_2)^\top (x_1 - x_2) + \frac{H}{2}\|x_1 - x_2\|^2.$$

Here, $\nabla f(x)$ denotes any subgradient of f at x. Strong convexity implies that the optimization problem $\min_{x \in X} f(x)$ has a unique solution. If $H = 0$ we simply say that the function is convex. We say a function g is H-strongly concave if $-g$ is H-strongly convex. Furthermore, we say a function $\mathcal{L}(x, y)$ is H-strongly convex-concave if for any fixed $y_0 \in Y$, the function $\mathcal{L}(x, y_0)$ is H-strongly convex in x, and for any fixed $x_0 \in X$, the function $\mathcal{L}(x_0, y)$ is H-strongly concave in y.

2.3. Saddle Points and Nash Equilibra

A pair (x^*, y^*) is called a saddle point for $\mathcal{L} : X \times Y \to \mathbb{R}$ if for any $x \in X$ and $y \in Y$, we have

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*). \tag{2}$$

It is well known that if $\mathcal{L}$ is convex-concave, and X and Y are convex and compact sets, there always exists at least one saddle point (see e.g. Boyd & Vandenberghe, 2004).

Moreover, if $\mathcal{L}$ is strongly convex-concave, the saddle point is unique.

A saddle point is also known as a Nash equilibrium for two-player zero-sum games (Nash, 1951). In a matrix game, the payoff function $\mathcal{L}(x,y) = x^\top Ay$ is bilinear, and therefore is convex-concave. The action spaces of the two players are $X = \Delta_{d_1}$ and $Y = \Delta_{d_2}$, which are convex and compact. As a result, there always exists a Nash equilibrium for any matrix game. The famous von Neumann minimax theorem states that $\min_{x\in\Delta_{d_1}} \max_{y\in\Delta_{d_2}} x^\top Ay = \max_{y\in\Delta_{d_2}} \min_{x\in\Delta_{d_1}} x^\top Ay$. If Player 1 chooses $x^* \in \arg\min_{x\in\Delta_{d_1}} \max_{y\in\Delta_{d_2}} x^\top Ay$ and Player 2 chooses $y^* \in \arg\max_{y\in\Delta_{d_2}} \min_{x\in\Delta_{d_1}} x^\top Ay$, the pair (x^*, y^*) is an equilibrium of the game (Morgenstern & Von Neumann, 1953).

2.4. Lipschitz Continuity

We say a function $f : X \to \mathbb{R}$ is G-Lipschitz continuous with respect to a norm $\|\cdot\|$ if for all $x, y \in X$ it holds that

$$|f(x) - f(y)| \leq G\|x - y\|$$

It is well known that the previous inequality holds if and only if

$$\|\nabla f(x)\|_* \leq G$$

for all $x \in X$, where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$ (Boyd & Vandenberghe, 2004; Shalev-Shwartz et al., 2012). Similarly, we say a function $\mathcal{L}(x,y)$ is G-Lipschitz continuous with respect to a norm $\|\cdot\|$ if

$$|\mathcal{L}(x_1,y_1) - \mathcal{L}(x_2,y_2)| \leq G\|[x_1;y_1] - [x_2;y_2]\|.$$

for any $x_1, x_2 \in X$ and any $y_1, y_2 \in Y$. Again, the previous inequality holds if and only if

$$\|[\nabla_x \mathcal{L}(x,y); \nabla_y \mathcal{L}(x,y)]\|_* \leq G$$

for all $x \in X, y \in Y$.

Lemma 1. *Consider a matrix A. If the absolute value of each entry of A is bounded by $c > 0$, then the function $\mathcal{L}(x,y) = x^\top Ay$ is $G_{\mathcal{L}}^{\|\cdot\|_2}$-Lipschitz continuous with respect to $\|\cdot\|_2$, where $G_{\mathcal{L}}^{\|\cdot\|_2} = \sqrt{c}\left(\sqrt{d_1} + \sqrt{d_2}\right)$. The function $\mathcal{L}$ is also $G_{\mathcal{L}}^{\|\cdot\|_1}$-Lipschitz continuous with respect to norm $\|\cdot\|_1$, where $G_{\mathcal{L}}^{\|\cdot\|_1} = c$.*

3. Challenges of the OMG Problem: An Impossibility Result

Recall that we defined the Online Matrix Games (OMG) problem in Section 1.1, where two players play a zero-sum games for T rounds. The sequence of payoff matrices $\{A_t\}_{t=1}^T$ is selected arbitrarily. In each round $t \in [T]$, both players choose their strategies before the payoff matrix A_t is revealed. The goal is to find strategies under which the players' average payoffs are close to the Nash Equilibrium of the game with payoff matrix $\sum_{t=1}^T A_t$.

Perhaps the most natural (albeit futile) approach to attack the OMG problem is to equip each of the players with a sublinear individual regret algorithm to generate a sequence of iterates $\{x_t, y_t\}_{t=1}^T$. We gave a few examples of Online Convex Optimization (OCO) algorithms that guarantee $O(\sqrt{T})$ regret in Section 1.3. However, if each player minimizes its individual regret greedily using OCO, this approach only implies that $\sum_{t=1}^T x_t^\top A_t y_t - \min_{x\in\Delta_X} \sum_{t=1}^T xA_t y_t = O(\sqrt{T})$, and $\max_{y\in\Delta_Y} \sum_{t=1}^T x_t^\top A_t y - \sum_{t=1}^T x_t^\top A_t y_t = O(\sqrt{T})$. Notice that the quantity $\min_{x\in X} \max_{y\in Y} \sum_{t=1}^T x^\top A_t y$ associated with the Nash Equilibrium in equation (1) does not even appear in these bounds. The reader familiar with saddle point computation may wonder how the so-called 'duality gap' (Bubeck et al., 2015): $\max_{y\in\Delta_Y} \sum_{t=1}^T x_t^\top A_t y - \min_{x\in\Delta_X} \sum_{t=1}^T xA_t y_t = O(\sqrt{T})$ relates to achieving sublinear NE regret. It is easy to see that the duality gap is the sum of individual regret of both players. In view of Theorem 1 we will see that NE regret and the duality gap are in some sense incompatible.

In this section we present a result that shows that there is no algorithm that *simultaneously* achieves sublinear NE regret and individual regret for both players. This implies that if both players individually use any existing algorithm from OCO they would inevitably fail to solve the OMG problem.

Theorem 1. *Consider any algorithm that selects a sequence of x_t, y_t pairs given the past payoff matrices $A_1, \ldots, A_{t-1}$. Consider the following three objectives:*

$$\left|\sum_{t=1}^T x_t^\top A_t y_t - \min_{x\in\Delta} \max_{y\in\Delta} \sum_{t=1}^T x^\top A_t y\right| \quad = \quad o(T), \quad (3)$$

$$\sum_{t=1}^T x_t^\top A_t y_t - \min_{x\in\Delta_X} \sum_{t=1}^T x^\top A_t y_t \quad = \quad o(T), \quad (4)$$

$$\max_{y\in\Delta_Y} \sum_{t=1}^T x_t^\top A_T y - \sum_{t=1}^T x_t^\top A_t y_t \quad = \quad o(T). \quad (5)$$

Then there exists an (adversarially-chosen) sequence $A_1, A_2, \ldots$ such that not all of (3), (4)*, and* (5)*, are true.*

A full proof of the result is shown in the Appendix, but here we give a sketch. The main idea is to construct two parallel scenarios, each with their own sequences of payoff matrices. The two scenarios will be identical for the first $T/2$ periods but are different for the rest of the horizon. In our particular construction, in both scenarios the players play the well known "matching-pennies" game for the first $T/2$ periods, then in first scenario they play a game with equal payoffs for all of their actions and in the second scenario they play a game where Player 1 is indifferent between its actions. One can show that if all three quantities in the statement of the theorem are $o(T)$ in the first scenario, then we prove that at least one of them is $\Omega(T)$ in the second one which yields the result. This suggests that the machinery for OCO, which minimizes individual regret, cannot be directly applied to the OMG problem.

4. Online Matrix Games: Full Information

4.1. Saddle Point Regularized Follow-the-Leader

In this section we propose an algorithm to solve the OMG problem in the full information setting. In fact, we will consider the algorithm in a slighly more general setting than the OMG problem, allowing the sequence of payoff functions to be specified by arbitrary convex-concave Lipschitz functions, and the action sets of Player 1 and Player 2 ($X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^n$ respectively) to be arbitrary convex compact sets.

Let the sequence of convex-concave functions be $\{\bar{\mathcal{L}}_t(x,y)\}_{t=1}^T$, which are $G_{\bar{\mathcal{L}}}$-Lipschitz with respect to some norm $\|\cdot\|$. We propose an algorithm called Saddle Point Regularized Follow the Leader (SP-RFTL), shown in Algorithm 1.

Algorithm 1 Saddle-Point Regularized-Follow-the-Leader (SP-RFTL)

```
input: $x_1 \in X$, $y_1 \in Y$, parameters: $\eta > 0$, strongly convex functions $R_X$, $R_Y$
for $t = 1, ...T$ do
    Play $(x_t, y_t)$
    Observe $\bar{\mathcal{L}}_t$
    $\mathcal{L}_t(x,y) \leftarrow \bar{\mathcal{L}}_t + \frac{1}{\eta}R_X(x) - \frac{1}{\eta}R_Y(y)$
    $x_{t+1} \leftarrow \arg\min_{x\in X}\max_{y\in Y}\sum_{\tau=1}^t \mathcal{L}_t(x,y)$
    $y_{t+1} \leftarrow \arg\max_{y\in Y}\min_{x\in X}\sum_{\tau=1}^t \mathcal{L}_t(x,y)$
end for
```

The regularizers R_X, R_Y are used as input for the algorithm. We will choose regularizers that are strongly convex with respect to norm $\|\cdot\|$, and G_{R_1} and G_{R_2} Lipschitz with respect to norm $\|\cdot\|$, which means that $\|\nabla R_X(x)\|_* \le G_{R_1}$ for all $x \in X$, and $\|\nabla R_Y(y)\|_* \le G_{R_2}$ for, all $y \in Y$. Finally, we assume $R_X(x) \ge 0$ for all $x \in X$ and $R_Y(y) \ge 0$ for all $y \in Y$.

The main difference between SP-RFTL and the well known Regularized Follow the Leader (RFTL) algorithm (Shalev-Shwartz & Singer, 2007; Abernethy et al., 2009) is that in SP-RFTL both players update jointly and play the saddle point of the sum of regularized games observed so far. In particular, they disregard their previous actions. In contrast, the updates for RFTL would be

$$x_{t+1}^{RFTL} \leftarrow \arg\min_{x\in X}\sum_{\tau=1}^t\left[\bar{\mathcal{L}}_\tau(x, y_\tau^{RFTL}) + \frac{1}{\eta}R_X(x)\right]$$

$$y_{t+1}^{RFTL} \leftarrow \arg\max_{y\in Y}\sum_{\tau=1}^t\left[\bar{\mathcal{L}}_\tau(x_\tau^{RFTL}, y) - \frac{1}{\eta}R_Y(y)\right]$$

for $t = 2, ..., T$, and x_1^{RFTL}, y_1^{RFTL} are chosen as to minimize $R_X(x)$ and $-R_Y(y)$ in their respective sets X, Y. It is easy to see that the sequence of iterates is in general not the same. In fact, in view of Theorem 1 we know that RFTL can not achieve sublinear NE regret when the sequence of functions is chosen arbitrarily. One last remark about the algorithm is that as $T \to \infty$ the last iterates (x_{T+1}, y_{T+1}) will converge to the set of NE of the average game $\frac{1}{T}\sum_{t=1}^T \bar{\mathcal{L}}_t$. To see this, observe that if $\eta = \sqrt{T}$ then $x_{T+1} \leftarrow \arg\min_{x\in X}\max_{y\in Y}\frac{1}{T}\sum_{t=1}^T\left[\bar{\mathcal{L}}_t(x,y)\right] + \frac{1}{\sqrt{T}}R_X(x) - \frac{1}{\sqrt{T}}R_Y(y)$ i.e. x_{T+1} solves the average problem where the regularization is vanishing, and a similar expression can be written for y_{T+1}. This is in contrast with many of the results mentioned in Section 1.3 where it is the *average* of the iterates which is an approximate equilibrium.

We have the following guarantee for SP-RFTL.

Theorem 2. *For $t = 1, ..., T$, let $\bar{\mathcal{L}}_t$ be $G_{\bar{\mathcal{L}}}$-Lipschitz with respect to norm $\|\cdot\|$. Let R_X, R_Y be strongly convex functions with respect to the same norm, let G_{R_X}, G_{R_Y} be the Lipschitz constants of R_X, R_Y with respect to the same norm. Let $\{(x_t, y_t)\}_{t=1}^T$ be the iterates generated by* SP-RFTL *when run on convex-concave functions $\{\bar{\mathcal{L}}_t(x,y)\}_{t=1}^T$. It holds that*

$$\left|\sum_{t=1}^T \bar{\mathcal{L}}_t(x_t,y_t) - \min_{x\in X}\max_{y\in Y}\sum_{t=1}^T \bar{\mathcal{L}}_t(x,y)\right|$$

$$\le 8\eta\left[G_{\bar{\mathcal{L}}} + \frac{1}{\eta}\max(G_{R_X}, G_{R_Y})\right]^2(1+\ln(T))$$

$$+ \frac{T}{\eta}\max_{y\in Y}R_Y(y) + \frac{T}{\eta}\max_{x\in X}R_X(x) \; = \; O\left(\sqrt{T\ln(T)}\right),$$

where the last equality follows by choosing $\eta = \frac{\sqrt{T}}{\ln(T)}$.

A formal proof of the theorem is provided in the Appendix and a sketch will be given shortly.

We note that the bound in Theorem 2 holds for general convex-concave functions, however the dependence on the dimension is hidden on the Lipschitz constants and the choice of regularizer. It is easy to check that if one chooses $\|\cdot\|_2^2$ as regularizer, and the functions $\{\mathcal{L}_t\}_{t=1}^T$ are G-Lipschitz continuous with respect to norm $\|\cdot\|_2^2$, then the NE regret bound will be $O(n\ln(T)\sqrt{T})$.

We now provide a sketch of the proof of Theorem 2. Define $\mathcal{L}_t(x,y) \triangleq \bar{\mathcal{L}}_t(x,y) + \frac{1}{\eta}R_X(x) - \frac{1}{\eta}R_Y(y)$. Notice that it is $\frac{1}{\eta}$-strongly convex in x with respect to norm $\|\cdot\|$ for all $y \in Y$ and $\frac{1}{\eta}$-strongly concave with respect to norm $\|\cdot\|$ for all $x \in X$. Additionally, notice that $\mathcal{L}_t$ is $G_{\mathcal{L}} \triangleq G_{\bar{\mathcal{L}}} + \frac{1}{\eta}(G_{R_X} + G_{R_Y})$-Lipschitz with respect to norm $\|\cdot\|$. Finally, notice that for $t = 1, ..., T$, all $x \in X$ and all $y \in Y$ it holds that

$$-\frac{1}{\eta}R_Y(y) \leq \mathcal{L}_t(x,y) - \bar{\mathcal{L}}_t(x,y) \leq \frac{1}{\eta}R_X(x) \qquad (6)$$

The following lemma shows that the value of the convex-concave games defined by $\sum_{t=1}^T \mathcal{L}_t$ and $\sum_{t=1}^T \bar{\mathcal{L}}_t$ are not too far from each other.

Lemma 2. *Let*

$$\begin{aligned}&\bar{x}_{T+1} \in \arg\min_{x\in X}\max_{y\in Y}\textstyle\sum_{t=1}^T \bar{\mathcal{L}}_t(x,y),\\ &\bar{y}_{T+1} \in \arg\max_{y\in Y}\min_{x\in X}\textstyle\sum_{t=1}^T \bar{\mathcal{L}}_t(x,y).\end{aligned}$$

It holds that

$$\begin{aligned}&-\frac{T}{\eta}R_Y(\bar{y}_{T+1})\\ &\leq \min_{x\in X}\max_{y\in Y}\sum_{t=1}^T \mathcal{L}_t(x,y) - \min_{x\in X}\max_{y\in Y}\sum_{t=1}^T \bar{\mathcal{L}}_t(x,y)\\ &\leq \frac{T}{\eta}R_X(\bar{x}_{T+1}).\end{aligned}$$

To prove the NE regret bound, we note that SP-RFTL is running a Follow-the-Leader scheme on functions $\{\mathcal{L}_{t=1}^T\}$ (Kalai & Vempala, 2002). With the next two lemmas one can show that the NE regret of the players relative to functions $\{\mathcal{L}\}_{t=1}^T$ is small.

Lemma 3. *Let $\{(x_t, y_t)\}_{t=1}^T$ be the iterates of* SP-RFTL. *It holds that*

$$\begin{aligned}&-G_{\mathcal{L}}\textstyle\sum_{t=1}^T \|x_t - x_{t+1}\|\\ &\leq \sum_{t=1}^T \mathcal{L}_t(x_{t+1}, y_{t+1}) - \min_{x\in X}\max_{y\in Y}\sum_{t=1}^T \mathcal{L}_t(x,y)\\ &\leq G_{\mathcal{L}}\textstyle\sum_{t=1}^T \|y_t - y_{t+1}\|.\end{aligned}$$

Lemma 4. *Let $\{(x_t, y_t)\}_{t=1}^T$ be the sequence of iterates generated by the algorithm. It holds that*

$$\begin{aligned}&\|x_t - x_{t+1}\| + \|y_t - y_{t+1}\|\\ &\leq \frac{4\eta}{t}\left[G_{\bar{\mathcal{L}}} + \frac{1}{\eta}\max(G_{R_X}, G_{R_Y})\right].\end{aligned}$$

Combining the NE regret bound obtained on functions $\{\mathcal{L}\}_{t=1}^T$ together with Lemma 2 and equation (6) yields the theorem.

4.2. Logarithmic Dependence on the Dimension of the Action Spaces

Previously, we analyzed the OMG problem by treating the payoff functions as general convex-concave functions and the action spaces as general convex compact sets. We explained that in general one should expect to achieve NE regret which depends linearly in the dimension of the problem. The goal in this section is to obtain sharper NE regret bounds that scale as $O(\ln(T)\sqrt{T}\ln(\max(d_1, d_2)))$ by exploiting the geometry of the decision sets Δ_X, Δ_Y and the bilinear structure of the payoff functions. This allows us to solve games which may have exponentially many actions, which often arise in combinatorial optimization settings.

The plan to obtain the desired NE regret bounds in this more restrictive setting is to use the negative entropy as a regularization function (which is strongly convex with respect to $\|\cdot\|_1$), that is $R_X(x) = \sum_{i=1}^{d_1} x_i \ln(x_i) + \ln(d_1)$ and $R_Y(y) = \sum_{i=1}^{d_2} y_i \ln(y_i) + \ln(d_2)$ where the extra logarithmic terms ensure R_X, R_Y are nonnegative everywhere in their respective simplexes. Unfortunately, the negative entropy is not Lipschitz over the simplex, so we can not leverage our result from Theorem 2. To deal with this challenge, we will restrict the new algorithm to play over a restricted simplex:[3]

$$\Delta_\theta = \{z \in \mathbb{R}^d : \|z\|_1 = 1, z_i \geq \theta, i = 1, ..., d\}. \qquad (7)$$

The tuning parameter $\theta \in [0, 1/d]$ used for the algorithm will be defined later in the analysis. (Notice that when $\theta > 1/d$, the set is empty.) We have the following result.

Lemma 5. *The function $R(x) \triangleq \sum_{i=1}^d x_i \ln(x_i)$ is G_R-Lipschitz continuous with respect to $\|\cdot\|_1$ over Δ_θ with $G_R = \max\{|\ln(\theta)|, 1\}$.*

The algorithm ONLINE-MATRIX-GAMES REGULARIZED-FOLLOW-THE-LEADER is an instantiation of SP-RFTL with a particular choice of regularization functions, which are nonegative and Lipschitz over the sets $\Delta_{X,\theta}$, $\Delta_{Y,\theta}$. With this, we can prove a NE regret bound for the OMG problem. For the remainder of the paper, the regularization functions will be set as follows:

$$\begin{aligned}R_X(x) &\triangleq \textstyle\sum_{i=1}^{d_1} x_i \ln(x_i) + \ln(d_1),\\ R_Y(y) &\triangleq \textstyle\sum_{i=1}^{d_2} y_i \ln(y_i) + \ln(d_2).\end{aligned}$$

We have the following guarantee for OMG-RFTL.

Theorem 3. *Let $\{A_t\}_{t=1}^T$ be an arbitrary sequence of matrices with entries bounded between $[-1, 1]$. Let $G_{\bar{\mathcal{L}}}$ be the Lipschitz constant (with respect to $\|\cdot\|_1$) of $\bar{\mathcal{L}}_t \triangleq x^\top A_t y$*

[3] We will also use the notation $\Delta_{X,\theta}$ and $\Delta_{Y,\theta}$ to mean the restricted simplex of Player 1 and 2, respectively

Algorithm 2 Online-Matrix-Games Regularized-Follow-the-Regularized-Leader (OMG-RFTL)

```
input: $x_1 \in \Delta_{X,\theta} \subset \mathbb{R}^{d_1}$, $y_1 \in \Delta_{Y,\theta} \subset \mathbb{R}^{d_2}$, parameters: $\eta > 0, \theta < \min\{\frac{1}{d_1}, \frac{1}{d_2}\}$.
for $t = 1, ...T$ do
    Play $(x_t, y_t)$, observe matrix $A_t$
    $\bar{\mathcal{L}}_t \leftarrow x^\top A_t y$
    $\mathcal{L}_t(x, y) \leftarrow \bar{\mathcal{L}}_t + \frac{1}{\eta} R_X(x) - \frac{1}{\eta} R_Y(y)$
    $x_{t+1} \leftarrow \arg\min_{x \in \Delta_{X,\theta}} \max_{y \in \Delta_{Y,\theta}} \sum_{\tau=1}^{t} \mathcal{L}_t(x, y)$
    $y_{t+1} \leftarrow \arg\max_{y \in \Delta_{Y,\theta}} \min_{x \in \Delta_{X,\theta}} \sum_{\tau=1}^{t} \mathcal{L}_t(x, y)$
end for
```

for $t = 1, ..., T$. Let $\{(x_t, y_t)\}_{t=1}^T$ be the iterates of OMG-RFTL*) and choose $\theta = e^{-\eta G_{\bar{\mathcal{L}}}} \leq \min\{\frac{1}{d_1}, \frac{1}{d_2}\}$ such that $\frac{|\ln(\theta)|}{\eta} = G_{\bar{\mathcal{L}}}$. Set $\eta = \frac{\sqrt{T}}{G_{\bar{\mathcal{L}}}}$. It holds that*

$$\begin{aligned}
&\left| \sum_{t=1}^{T} x_t^\top A_t y_t - \min_{x \in \Delta} \max_{y \in \Delta} \sum_{t=1}^{T} x^\top A_t y \right| \\
&\leq 32 G_{\bar{\mathcal{L}}} \sqrt{T}(1 + \ln(T)) + 2\sqrt{T} \max\{\ln d_1, \ln d_2\} + \\
&\quad 2 \max\{d_1, d_2\} G_{\bar{\mathcal{L}}} T e^{-\sqrt{T}} \\
&= O\left(\ln(T)\sqrt{T} + \sqrt{T} \max\{\ln d_1, \ln d_2\}\right) + \\
&\quad o(1) \max\{d_1, d_2\}.
\end{aligned}$$

A full proof of the theorem can be found in the Appendix. We now give a sketch of the proof. Since the algorithm selects actions over the restricted simplex, we must quantify the potential loss in the NE regret bound imposed by this restriction. The next two lemmas make this precise.

Lemma 6. *Let $z^* \in \Delta \subset \mathbb{R}^d$ define $z_p^* \triangleq \arg\min_{z \in \Delta_\theta} \|z - z^*\|_1$, with $\theta \leq \frac{1}{d}$. Notice z_p^* is unique since it is a projection. It holds that $\|z_p^* - z^*\|_1 \leq 2\theta(d-1)$.*

Lemma 7. *Let $\{\bar{\mathcal{L}}_t(x, y)\}_{t=1}^T$ be an arbitrary sequence of convex-concave functions, $\bar{\mathcal{L}}_t : \Delta_X \times \Delta_Y \to \mathbb{R}$, that are $G_{\bar{\mathcal{L}}}$-Lipschitz with respect to $\|\cdot\|_1$. With $\Delta_X \subseteq \mathbb{R}^{d_1}$, and $\Delta_Y \subseteq \mathbb{R}^{d_2}$. It holds that*

$$\begin{aligned}
&- G_{\bar{\mathcal{L}}} T \|x_p^* - x^*\|_1 \\
&\leq \min_{x \in \Delta} \max_{y \in \Delta} \sum_{t=1}^{T} \bar{\mathcal{L}}_t(x, y) - \min_{x \in \Delta_\theta} \max_{y \in \Delta_\theta} \sum_{t=1}^{T} \bar{\mathcal{L}}_t(x, y) \\
&\leq G_{\bar{\mathcal{L}}} T \|y_p^* - y^*\|_1.
\end{aligned}$$

Combining the previous two lemmas and Theorem 2, one can show the NE regret bound for OMG-RFTL holds.

5. Online Matrix Games: Bandit Feedback

In this section we focus on the OMG problem under bandit feedback. In this setting, the players observe in every round only the payoff corresponding to the chosen actions. If Player 1 chooses action i, Player 2 chooses action j, and the payoff matrix at that time step is A_t, then the players observe only $(A_t)_{ij}$ instead of the full matrix A_t. The limited feedback makes the problem significantly more challenging than the full information one: the players must find a way to *exploit* (use all previous information to try to play a Nash Equilibrium) and *explore* (try to estimate A_t in every round). This problem resembles that of Online Bandit Optimization (Flaxman et al., 2005; Auer et al., 1995; Bubeck et al., 2016; Hazan & Li, 2016), while the main difference is that with one function evaluation we must estimate a matrix A_t instead of the gradients $\nabla_x \mathcal{L}_t(x, y)$ and $\nabla_y \mathcal{L}_t(x, y)$ where $\mathcal{L}_t = x^\top A_t y$.

Before proceeding we establish some useful notation. For $i = 1, ..., d$, let $e_i \in \mathbb{R}^d$ be the collection of standard unit vectors i.e. e_i is the vector that has a 1 in the i-th entry and 0 in the rest. Let $e_{x,t}$ be the standard unit vector corresponding to the decision made by Player 1 for round t, define $e_{y,t}$ similarly. Notice that under bandit feedback, in round t both players only observe the quantity $e_{x,t}^\top A_t e_{y,t}$.

5.1. A One-Point Estimate for $\mathcal{L}(x, y) = x^\top A y$

As explained previously, in each round t the players must estimate A_t by observing only one of its entries. To this end, we allow the players to share with each other their decisions and to randomize *jointly* (a similar assumption is used to define correlated equilibria in zero-sum games, see Aumann (1987)). The following result shows how to build a random estimate of A by observing only one of its entries.

Theorem 4. *Let $x \in \Delta_{X,\delta}, y \in \Delta_{Y,\delta}$ with $d_1, d_2 \geq 2$ and $\delta > 0$. Sample $i' \sim x, j' \sim y$. Let $\hat{A}$ be the $d_1 \times d_2$ matrix with $\hat{A}_{i,j} = 0$ for all i, j such that $i \neq i'$ and $j \neq j'$ and $\hat{A}_{i',j'} = \frac{A_{i',j'}}{x(i')y(j')}$. It holds that*

$$\mathbb{E}_{i' \sim x, j' \sim y}[\hat{A}] = A.$$

5.2. Bandit Online Matrix Games RFTL

We now present an algorithm that ensures sublinear (i.e. $o(T)$) NE regret under bandit feedback for the OMG problem that holds against an adaptive adversary. By adaptive adversary, we mean that the payoff matrices A_t can depend on the players' actions up to time $t-1$; in particular, we assume the adversary does not observe the actions chosen by the players for time period t when choosing A_t. We consider an algorithm that runs OMG-RFTL on a sequence of functions $\hat{\mathcal{L}}_t \triangleq x^\top \hat{A}_t y$, where $\hat{A}_t$ is the unbiased one-point estimate of A_t derived in Theorem 4. Recall that the iterates of OMG-RFTL algorithm are distributions over the possible actions of both players. In order to generate the estimate $\hat{A}_t$, both players will sample an action from their distributions and weigh their observation with the inverse probability of obtaining that observation.

Algorithm 3 Bandit Online-Matrix-Games Regularized-Follow-the-Leader (BANDIT-OMG-RFTL)

```
**input:** $x_1 \in \Delta_{X,\delta} \subset \mathbb{R}^{d_1}$, $y_1 \in \Delta_{Y,\delta} \subset \mathbb{R}^{d_2}$, parameters: $\eta > 0$, $0 < \delta < \min\{\frac{1}{d_1}, \frac{1}{d_2}\}$.
**for** $t = 1, ...T$ **do**
    Sample independently $e_{x,t} \sim \tilde{x}_t$ and $e_{y,t} \sim \tilde{y}_t$
    Observe $e_{x,t}^\top A_t e_{y,t}$
    Build $\hat{A}_t$ as in Theorem 4 using $e_{x,t}^\top A_t e_{y,t}, x_t, y_t$
    $\hat{\mathcal{L}}_t \leftarrow x^\top \hat{A}_t y$
    $\mathcal{L}_t(x,y) \leftarrow \hat{\mathcal{L}}_t + \frac{1}{\eta} R_X(x) - \frac{1}{\eta} R_Y(y)$
    $x_{t+1} \leftarrow \arg\min_{x \in \Delta_{X,\theta}} \max_{y \in \Delta_{Y,\theta}} \sum_{\tau=1}^t \mathcal{L}_t(x,y)$
    $y_{t+1} \leftarrow \arg\max_{y \in \Delta_{Y,\theta}} \min_{x \in \Delta_{X,\theta}} \sum_{\tau=1}^t \mathcal{L}_t(x,y)$
**end for**
```

We have the following guarantee for BANDIT-OMG-RFTL.

Theorem 5. *Let $\{A_t\}_{t=1}^T$ be any sequence of payoff matrices chosen by an adaptive adversary. Let $\{e_{x,t}, e_{y,t}\}_{t=1}^T$ be the iterates generated by* BANDIT-OMG-FTRL. *Setting $\delta = \frac{1}{T^{1/6}}$, $\eta = T^{1/6}$ ensures*

$$\left| \mathbb{E}\left[\sum_{t=1}^T e_{x,t}^\top A_t e_{y,t} - \min_{x \in X} \max_{y \in Y} \sum_{t=1}^T x^\top A_t y \right] \right|$$
$$\le O((d_1 + d_2)\ln(T) T^{5/6})$$

where the expectation is taken with respect to randomization in the algorithm.

We now give a sketch of the proof. The total payoff given to each of the players is given by $\sum_{t=1}^T e_{x,t}^\top A_t e_{y,t}$ so we must relate this quantity to the iterates $\{x_t, y_t\}_{t=1}^T$ of OMG-RFTL when run on sequence of matrices $\{\hat{A}_t\}_{t=1}^T$. The following two lemmas will allow us to do so.

Lemma 8. *Let $\{e_{x,t}, e_{y,t}\}_{t=1}^T$ be the sequence of iterates generated by* BANDIT-OMG-RFTL. *It holds that*

$$\mathbb{E}\left[\textstyle\sum_{t=1}^T e_{x,t}^\top A_t e_{y,t}\right] = \mathbb{E}\left[\textstyle\sum_{t=1}^T x_t^\top A_t y_t\right],$$

where the expectation is taken with respect to the internal randomness of the algorithm.

Lemma 9. *It holds that*

$$\mathbb{E}\left[\textstyle\sum_{t=1}^T x_t^\top \hat{A}_t y_t\right] = \mathbb{E}\left[\textstyle\sum_{t=1}^T x_t^\top A_t y_t\right],$$

where the expectation is with respect to all the internal randomness of the algorithm.

We will then bound the difference between the comparator term $\min_{x \in \Delta} \max_{y \in \Delta} \sum_{t=1}^T x^\top A_t y$ and the comparator term Theorem 3 gives us by running OMG-RFTL on functions $\{\hat{\mathcal{L}}\}_{t=1}^T$, $\min_{x \in \Delta} \max_{y \in \Delta} \sum_{t=1}^T x^\top \hat{A}_t y$. Special care must be taken to ensure this difference holds even against an adaptive adversary. To this end, we use the next two lemmas; in particular, the proof of Lemma 11 relies heavily on Theorem 4.

Lemma 10. *With probability 1 it holds that*

$$\left| \min_{x \in \Delta_X^\delta} \max_{y \in \Delta_Y^\delta} \sum_{t=1}^T x^\top A_t y - \min_{x \in \Delta_X^\delta} \max_{y \in \Delta_Y^\delta} \sum_{t=1}^T x^\top \hat{A}_t y \right|$$
$$\le \left\| \textstyle\sum_{t=1}^T A_t y - \hat{A}_t y \right\|_2.$$

Lemma 11. *It holds that*

$$\mathbb{E}\left[\left\| \sum_{t=1}^T A_t y - \hat{A}_t y \right\|_2 \right] \le \frac{2\sqrt{T}\min(d_1, d_2)}{\delta^2},$$

where the expectation is taken with respect to the internal randomness of the algorithm.

The proof of Theorem 5 follows by combining Lemmas 8 through 11, with careful choice of tuning parameters.

6. Conclusion

In this paper, we considered the Online Matrix Games problem, where two players interact in a sequence of zero-sum games with arbitrarily changing payoff matrices. The goal for both players is to achieve small Nash Equilibrium (NE) regret, that is, the players want to ensure their average payoffs over T rounds are close to those in the NE of the mean payoff matrix in hindsight. While it is known that standard Online Convex Optimization algorithms such as Online Gradient Descent can be used to find approximate equilibria in *static* zero-sum games, our impossibility result shows that no algorithm for online convex optimization can achieve sublinear Nash Equilibrium regret ($o(T)$) when the sequence of payoffs are chosen arbitrarily. We then design and analyze algorithms that achieve sublinear NE regret for the Online Matrix Games problem, under both full information feedback and bandit feeback settings. In the full information case, the performance of the algorithm is optimal with respect to the number of rounds (up to logarithm factors) and depends logarithmically on the number of actions of each player. For the bandit feedback setting, we provide an algorithm with sublinear NE regret using a one-point matrix estimate.

7. Acknowledgements

We thank the reviewer who noticed an issue with our initial one-point estimate estimate of matrix A_t and his/her suggestion on how to fix it.

References

Abernethy, J., Lai, K. A., Levy, K. Y., and Wang, J.-K. Faster rates for convex-concave games. *arXiv preprint arXiv:1805.06792*, 2018.

Abernethy, J. D. and Wang, J.-K. On Frank-Wolfe and equilibrium computation. In *Advances in Neural Information Processing Systems*, pp. 6587–6596, 2017.

Abernethy, J. D., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. *In Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2009.

Adler, I. The equivalence of linear programs and zero-sum games. *International Journal of Game Theory*, 42(1): 165–177, 2013.

Agrawal, S., Wang, Z., and Ye, Y. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *focs*, pp. 322. IEEE, 1995.

Aumann, R. J. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pp. 1–18, 1987.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 207–216. IEEE, 2013.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. The mechanics of n-player differentiable games. *arXiv preprint arXiv:1802.05642*, 2018.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*, volume 28. Princeton University Press, 2009.

Bowling, M. Convergence and no-regret in multiagent learning. In *Advances in neural information processing systems*, pp. 209–216, 2005.

Bowling, M. and Veloso, M. Convergence of gradient dynamics with a variable learning rate. In *ICML*, pp. 27–34, 2001.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Bubeck, S., Eldan, R., and Lee, Y. T. Kernel-based methods for bandit convex optimization. *arXiv preprint arXiv:1607.03084*, 2016.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

Conitzer, V. and Sandholm, T. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, 67(1-2):23–43, 2007.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Ferreira, K., Simchi-Levi, D., and Wang, H. Online network revenue management using thompson sampling. *Operations Research*, 2018. (forthcoming).

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394. Society for Industrial and Applied Mathematics, 2005.

Freund, Y. and Schapire, R. E. Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on Computational learning theory*, pp. 325–332. ACM, 1996.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Granichin, O. Stochastic approximation with input perturbation under dependent observation noises. - . *1. . .* , 4: 27–31, 1989.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Hazan, E. and Li, Y. An optimal algorithm for bandit convex optimization. *arXiv preprint arXiv:1603.04350*, 2016.

Ho-Nguyen, N. and Kılınç-Karzan, F. The role of flexibility in structure-based acceleration for online convex optimization. Technical report, Carnegie Mellon University, 2016. Technical report, http://www. optimization-online. org/DB_HTML/2016/08/5571.html.

Immorlica, N., Sankararaman, K. A., Schapire, R., and Slivkins, A. Adversarial bandits with knapsacks. *arXiv preprint arXiv:1811.11881*, 2018.

Kalai, A. and Vempala, S. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, 3 (Nov):423–440, 2002.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 157–163. Elsevier, 1994.

Morgenstern, O. and Von Neumann, J. *Theory of games and economic behavior*. Princeton university press, 1953.

Nash, J. Non-cooperative games. *Annals of mathematics*, pp. 286–295, 1951.

Shalev-Shwartz, S. and Singer, Y. *Online learning: Theory, algorithms, and applications*. Citeseer, 2007.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Singh, S., Kearns, M., and Mansour, Y. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pp. 541–548. Morgan Kaufmann Publishers Inc., 2000.

Spall, J. C. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1): 109–112, 1997.

Von Neumann, J. Die zerlegung eines intervalles in abzählbar viele kongruente teilmengen. *Fundamenta Mathematicae*, 1(11):230–238, 1928.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 928–936, 2003.