Online Adversarial Zero-Sum Games

Omid Sadeghi
Department of Electrical and Computer Engineering
University of Washington
Seattle, WA 98195
omids@uw.edu

Abstract

In the classical learning setting in zero-sum games, it is assumed that the underlying payoff (cost) function is fixed. In contrast, I consider the following online zero-sum game: The domains X, Y and the time horizon T are given offline. At each round $t \in [T]$, each player chooses an action $x_t \in X$ and $y_t \in Y$ and upon committing to these actions, a function $\mathcal{L}_t(x, y);\ x \in X, y \in Y$ is revealed, the first player incurs the cost $\mathcal{L}_t(x_t, y_t)$ and the second player obtains the reward $\mathcal{L}_t(x_t, y_t)$. [1, 2] studied this problem for the special case where both X and Y are the simplex and for all $t \in [T]$, the function $\mathcal{L}_t(x, y) = x^T A_t y$ is bi-linear. This setting is called Online Matrix Games. Moreover, [3, 4, 5] considered the more general Online Saddle Point problem framework in which X and Y are convex and compact and the functions $\{\mathcal{L}_t(x, y)\}_{t=1}^T$ are (strongly) convex in $x \in X$ and (strongly) concave in $y \in Y$. While the goal in all the previous works is to minimize the regret (i.e., sub-linear regret), this problem has been studied under various notions of regret and different feedback settings. In this project, I focus on the case where $\{\mathcal{L}_t(x, y)\}_{t=1}^T$ is chosen adversarially, and I explore the different problem settings and regret metrics, and study the proposed algorithms and their performance guarantees.

1 Introduction

The Online Convex Optimization (OCO) framework covers a large number of problems in which information is revealed incrementally (i.e., *online*) and irrevocable decisions should be made at each step in face of uncertainty about the future arriving information [6, 7]. Such problems could be formulated as a repeated game between the decision maker (i.e., the learner) and the adversary (i.e., the nature or environment). In the standard OCO problem, a convex domain X and a finite time horizon T are given. At each iteration $t \in [T]$, the learner chooses an action $x_t \in X$ based on past information from time steps $1, \dots, t-1$ and then, a convex loss function f_t is revealed and the learner incurs the cost $f_t(x_t)$ for her selected action. In the non-stochastic (adversarial) model, no assumptions are made on the sequence of arriving loss functions except their boundedness. As time goes by, the learner aims to observe the past and make better decisions to maximize the overall reward. The performance of an online algorithm for the OCO problem is typically measured through the *regret* of the algorithm which characterizes the difference between the total loss incurred by the algorithm and that of the best fixed benchmark action with hindsight information [6, 8, 9, 10], i.e., $\frac{1}{T}\sum_{t=1}^T f_t(x_t) - \min_{x \in X} \frac{1}{T}\sum_{t=1}^T f_t(x)$. This problem is very well-studied and several algorithms including Online Gradient Descent, Regularized Follow the Leader and Perturbed Follow the Leader achieve the provably optimal $\mathcal{O}(\sqrt{T})$ regret bound.

The OCO problem involves only a single individual playing against nature. However, in many applications, two players interact in a zero-sum game repeatedly where the payoff functions evolve arbitrarily over T rounds. Zero-sum games have numerous applications in economics and are important to understanding linear programming duality, convex optimization, robust optimization and differential privacy. I consider a natural extension of the OCO problem to the two-player setting

Preprint. Under review.

where domains X, Y and the time horizon T are given. At each round $t \in [T]$, player 1 and player 2 simultaneously select $x_t \in X$ and $y_t \in Y$ and upon committing to these actions, the payoff function $\mathcal{L}_t(x, y); x \in X, y \in Y$ is revealed and the two players receive $\mathcal{L}_t(x_t, y_t)$ as the loss and utility respectively.

1.1 Related work

Saddle Point Optimization: The study of saddle point problems in the context of online learning is extremely limited. I will mainly discuss [1, 2, 3, 4] in this project. Prior to these recent works, [11] provided a detailed overview of online learning for static two-player zero-sum games where $\mathcal{L}_t(x, y) = \mathcal{L}(x, y)$ for all $t \in [T]$ and $\mathcal{L}$ is convex in x and concave in y. For this setting, [12] showed that if both players employ a no-regret algorithm to minimize their individual regret, then their average of actions $(\bar{x}, \bar{y})$ satisfy $|\mathcal{L}(\bar{x}, \bar{y}) - \mathcal{L}(x^*, y^*)| \to 0$ as $T \to \infty$, where (x^*, y^*) is a Nash equilibrium. [13, 14] proposed a modification of the gradient ascent algorithm, called WoLF (Win or Learn Fast) whose iterates converge to the Nash equilibrium. This result was later generalized to multi-player non-zero-sum static games by [15] and their algorithm called GIGA-WoLF. As mentioned before, unlike the focus of this project, all the aforementioned works focus on repeated games with static payoff functions.
Online Convex Optimization (OCO) with Constraints: [16] introduced a special case of online saddle point problems to handle difficult convex constraints in the OCO problem. The difficult constraints (i.e., sets which it is computationally expensive to project onto them) $\{g^{(i)}(x) \leq 0\}_{i=1}^m$ are embedded into each loss function $f_t(x)$ through forming the corresponding Lagrangian $f_t(x) + \sum_{i=1}^m y^{(i)} g^{(i)}(x)$ which is convex in x and concave in y. At each round $t \in [T]$, both primal and dual variables x, y are updated and the goal is to obtain sub-linear regret and sub-linear constraint violation $\sum_{t=1}^T g^{(i)}(x_t)$ for all $i \in [m]$. This idea of modeling online constrained problems as online saddle point problems has been extensively studied later on as well [17, 18].
Adversarial Bandits with Knapsacks: [19] introduced a generalization of the standard Multi Armed Bandits problem in which there are a finite set of arms available and upon pulling each of the arms, an arbitrarily chosen reward is received while consuming some resources. The overall goal is to maximize the cumulative reward without exceeding a pre-specified total available budget. Similar to OCO with constraints, this problem could be modeled as an online zero-sum game through using the Lagrangian function and [19] managed to obtain sub-linear regret bounds for this problem without violating the budget constraints.

1.2 Outline

In Section 2, I will first introduce some notations and definitions that will be used in later sections. The problem framework is formally described in Section 3 and some motivating applications are provided in Section 3.2. In Section 3.1, I define the two performance metrics commonly used in the literature for this problem and then, I will discuss a simple counterexample in Section 3.1.1 showing that these two metrics are incompatible. In Section 4, I will discuss the algorithms for the Online Saddle Point (OSP) problem in the full information setting and provide their performance guarantees. For the special case that payoff functions are bi-linear (Online Matrix Games (OMG)), I will discuss an algorithm for the bandit feedback setting and its theoretical guarantees in Section 5. Finally, I will close with a summary of discussed results along with potential future research directions in Section 6.

2 Preliminaries

Notations. We use $[T]$ to denote the set $\{1, 2, \ldots, T\}$. For a vector $x \in \mathbb{R}^n$, we use x_i to denote the i-th entry of x and for a matrix A, A_{ij} denotes the entry in the i-th row and j-th column. $[x; y]$ denotes the concatenation of two vectors x, y as a column vector. The inner product of two vectors $x, y \in \mathbb{R}^n$ is denoted by either $\langle x, y \rangle$ or $x^T y$. Also, $B_R(z_1, z_2) = R(z_1) - R(z_2) - \langle \nabla R(z_2), z_1 - z_2 \rangle$ is the Bregman Divergence with respect to the function R.
Convexity and Strong Convexity. A function $f : X \to \mathbb{R}$ is called μ-strongly convex with respect to a norm $\|\cdot\|$ if for any $x_1, x_2 \in X$, the following holds:

$$f(x_1) \geq f(x_2) + \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{\mu}{2} \|x_1 - x_2\|^2,$$

where $\nabla f(x)$ denotes any sub-gradient of f at x. If $\mu = 0$, f is called convex. g is μ-strongly concave (concave) if $-g$ is μ-strongly convex (convex).
Saddle Points and Nash Equilibrium. For a function $\mathcal{L} : X \times Y \to \mathbb{R}$, a pair (x^*, y^*) is called a saddle point if for any $x \in X$ and $y \in Y$, we have the following:

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

It is known that if $\mathcal{L}$ is convex in x and concave in y, and X, Y are convex and compact sets, at least one saddle point always exists. Furthermore, if $\mathcal{L}$ is strongly convex in x and strongly concave in y, the saddle point is unique. A saddle point is also called a Nash equilibrium in the context of a two-player zero-sum game whose payoff function is characterized by $\mathcal{L}$.
Lipschitz Continuity. A function $f : X \to \mathbb{R}$ is called G-Lipschitz with respect to a norm $\|\cdot\|$ if for all $x_1, x_2 \in X$, we have:

$$|f(x_1) - f(x_2)| \leq G\|x_1 - x_2\|,$$

or equivalently, $\|\nabla f(x)\|_* \leq G$ holds for all $x \in X$. Similarly, $\mathcal{L}(x, y)$ is G-Lipschitz with respect to a norm $\|\cdot\|$ if for any $x_1, x_2 \in X$ and $y_1, y_2 \in Y$, we have $|\mathcal{L}(x_1, y_1) - \mathcal{L}(x_2, y_2)| \leq G\|[x_1; y_1] - [x_2; y_2]\|$ or equivalently, $\|[\nabla_x \mathcal{L}(x, y); \nabla_y \mathcal{L}(x, y)]\|_* \leq G$ holds.

3 Online Saddle Point (OSP) problem

The framework of Online Saddle Point (OSP) problem is as follows:

- Convex and compact domains X, Y and the time horizon T are given.
- Two players interact in a repeated zero-sum game through T rounds.
- At each round $t \in [T]$, player 1 and player 2 simultaneously select from domains X and Y respectively.
- Upon committing to these actions, the arbitrarily chosen function $\mathcal{L}_t(x, y)$ is revealed which is convex in x, concave in y and G-Lipschitz. Player 1 incurs the cost $\mathcal{L}_t(x_t, y_t)$ and player 2 obtains the reward $\mathcal{L}_t(x_t, y_t)$.

The special case where $\mathcal{L}_t(x, y) = x^T A_t y$ for all $t \in [T]$ and X, Y are the simplex is called *Online Matrix Games (OMG)*.

3.1 Performance metric

In order to quantify the performance of the proposed algorithms for this problem, two notions of *regret* have been used in the literature.
Nash Equilibrium (NE) regret. This measure was introduced by [1, 2, 4] and it quantifies the difference between the average payoff of the players and that of the Nash equilibrium of the average payoff functions:

$$\text{NE-Regret}(T) = |\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(x_t, y_t) - \min_{x \in X} \max_{y \in Y} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_t(x, y)|.$$

Weighted Online Saddle Point (WOSP) gap. This metric was introduced by [3] and is the sum of weighted individual regret of the players, i.e., WOSP-gap(T) = WI1-Regret(T) + WI2-Regret(T) and

$$\text{WI1-Regret}(T) = \sum_{t=1}^{T} \theta_t \mathcal{L}_t(x_t, y_t) - \min_{x \in X} \sum_{t=1}^{T} \theta_t \mathcal{L}_t(x, y_t),$$

$$\text{WI2-Regret}(T) = \max_{y \in Y} \sum_{t=1}^{T} \theta_t \mathcal{L}_t(x_t, y) - \sum_{t=1}^{T} \theta_t \mathcal{L}_t(x_t, y_t),$$

where $\theta_t \geq 0 \; \forall t \in [T]$ and $1^T\theta = 1$. Note that if we set $\theta_t = \frac{1}{T}$ for all $t \in [T]$, WI1-Regret(T) and WI2-Regret(T) are simply the standard notion of individual regret used in the OCO literature. Weighted regret enables us to model situations in which the decisions made throughout the $t \in [T]$ rounds have varying importance that are captured by the size of their corresponding weight θ_t.

3.1.1 Incompatibility of performance metrics

Theorem 1. *Consider any algorithm for the OMG problem that for all $t \in [T]$, selects a sequence of (x_t, y_t) pairs given the past payoff matrices $A_1, \ldots, A_{t-1}$. Let $\theta_t = \frac{1}{T}\ \forall t \in [T]$. Then, there exists an adversarially chosen sequence $A_1, \ldots, A_T$ such that not all NE-Regret$(T) = o(1)$, WI1-Regret$(T) = o(1)$ and WI2-Regret$(T) = o(1)$ are true.*

Proof. Assume that an algorithm exists such that for all sequences of matrices $\{A_t\}_{t=1}^T$ with bounded entries in $[-1, 1]$, all of the following hold:

$$|\sum_{t=1}^{T} x_t^T A_t y_t - \min_{x\in\Delta_X} \max_{y\in\Delta_Y} \sum_{t=1}^{T} x^T A_t y| \leq o(T), \tag{1}$$

$$\sum_{t=1}^{T} x_t^T A_t y_t - \min_{x\in\Delta_X} \sum_{t=1}^{T} x^T A_t y_t \leq o(T), \tag{2}$$

$$\max_{y\in\Delta_Y} \sum_{t=1}^{T} x_t^T A_t y - \sum_{t=1}^{T} x_t^T A_t y_t \leq o(T). \tag{3}$$

The proof is based on constructing two sequences of matrices $\{A_t\}_{t=1}^T$ for which all the three guarantees hold and lead that to a contradiction. Let T be divisible by 2 and consider the following two scenarios:

Scenario 1: $A_t = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\ 1 \leq t \leq \frac{T}{2}$ & $A_t = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\ \frac{T}{2} < t \leq T.$

Scenario 2: $A_t = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\ 1 \leq t \leq \frac{T}{2}$ & $A_t = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}\ \frac{T}{2} < t \leq T.$

In both scenarios, we have $\min_{x\in\Delta_X} \max_{y\in\Delta_Y} \sum_{t=1}^{T} x^T A_t y = 0$ because:

$$\min_{x\in\Delta_X} \max_{y\in\Delta_Y} \sum_{t=1}^{T} x^T A_t y = \min_{x\in\Delta_X} \max_{y\in\Delta_Y} \frac{T}{2} x^T \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} y = \frac{T}{2} \min_{x\in\Delta_X} \max\{x_1 - x_2, -x_1 + x_2\} = 0,$$

$$\min_{x\in\Delta_X} \max_{y\in\Delta_Y} \sum_{t=1}^{T} x^T A_t y = \min_{x\in\Delta_X} \max_{y\in\Delta_Y} T x^T \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} y = T \min_{x\in\Delta_X} \max\{x_1, -x_1\} = 0,$$

in scenarios 1 and 2 respectively. Let $x = [\alpha; 1-\alpha]$ and $y = [\beta; 1-\beta]$ for some $0 \leq \alpha, \beta \leq 1$. By inequalities 1 and 3, $\max_{y\in\Delta_Y} \sum_{t=1}^{T} x_t^T A_t y - \min_{x\in\Delta_X} \max_{y\in\Delta_Y} \sum_{t=1}^{T} x^T A_t y \leq o(T)$ holds. Therefore, in scenario 1 and scenario 2 respectively, we have:

$$\max_{y\in\Delta_Y} \sum_{t=1}^{T} x_t^T A_t y = \max_{\beta\in[0,1]} \sum_{t=1}^{\frac{T}{2}} \big(4\alpha_t\beta - 2\beta + 1 - 2\alpha_t\big) = \sum_{t=1}^{\frac{T}{2}} |2\alpha_t - 1| \leq o(T),$$

$$\max_{y\in\Delta_Y} \sum_{t=1}^{T} x_t^T A_t y = \max_{\beta\in[0,1]} \sum_{t=1}^{\frac{T}{2}} \big(4\alpha_t\beta - 2\beta + 1 - 2\alpha_t\big) + \frac{T}{2}(2\beta - 1) = \sum_{t=1}^{\frac{T}{2}} |2\alpha_t - 1| + \frac{T}{2} \leq o(T).$$

Therefore, $\sum_{t=1}^{\frac{T}{2}} (2\alpha_t - 1) + \frac{T}{2} \leq o(T)$ and $\sum_{t=1}^{\frac{T}{2}} (-2\alpha_t + 1) \leq o(T)$ should hold simultaneously leading to $\frac{T}{2} \leq o(T)$ which is a contradiction. □

3.2 Motivating examples

Minimizing individual regrets of both players as the performance metric has been widely studied in the literature. However, the choice of benchmark in the NE-regret might look surprising at first. Nonetheless, there are a number of interesting applications in which the comparator term $\min_{x\in X} \max_{y\in Y} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(x, y)$ arises naturally and therefore, the NE-regret is a better notion of regret for such problems. A number of these examples are listed below:

- Online Packing Problems: An online optimization problem in which player 1 chooses the primal variable and player 2 selects the dual variable (corresponding to the packing

constraints) at each round. Using Lagrangian duality, this problem can be modeled as a zero-sum game with the Lagrangian function $\mathcal{L}_t$ as the payoff function and the benchmark $\min_{x\in X}\max_{y\in Y}\sum_{t=1}^{T}\mathcal{L}_t(x,y)$ is the optimal value of the optimization problem.

- Adversarial Bandits with Knapsacks: An extension of the classical Multi Armed Bandits (MAB) problem in which there are a finite set of arms and upon pulling an arm, the reward and resource consumption of the arm are revealed. The goal is to maximize the overall reward while honoring the resource constraints. Similar to online packing problems, Lagrangian duality could be used to model the problem as a zero-sum game between player 1 and player 2 who choose the primal and dual variables respectively.
- Generative Adversarial Networks (GANs): A zero-sum game where the decision-maker trains the generator and the discriminator to find a Nash equilibrium. In this example, both players desire to update jointly and therefore, NE-regret is a more suitable notion of performance metric.

4 Online Saddle Point (OSP) problem: Full information

As I mentioned in Section 3.1.1, the two widely-used notions of regret in the literature for the OSP problem are incompatible, and therefore, the proposed algorithms for this problem focus on optimizing one of the two performance metrics without having any guarantees for the other. Firstly, I will discuss the Online Generalized Mirror Descent algorithm proposed by [3] which aims to obtain sub-linear WOSP-gap bounds. This algorithm is a simple application of the well-known Online Mirror Descent algorithm to the OSP problem. In particular, player 1 and player 2 each use the Online Mirror Descent algorithm equipped with strongly convex regularizer functions R_X and R_Y respectively in order to minimize their individual regrets. The main difference with the standard OCO literature is the fact that the regret weights $\{\theta_t\}_{t=1}^{T}$ appear as the coefficients for the gradient of payoff functions in the update rule of the algorithm.

Algorithm Online Generalized Mirror Descent

Input: Positive step sizes $\{\gamma_t\}_{t=1}^{T}$ and regret weights $\{\theta_t\}_{t=1}^{T}$, $\beta_X, \beta_Y > 0$, 1-strongly convex functions R_X and R_Y.
Output: $\{(x_t, y_t) : 1 \le t \le T\}$.
Set $x_1 = \arg\min_{x\in X} R_X(x)$, $y_1 = \arg\max_{y\in Y} R_Y(y)$.
for $t = 1$ **to** T **do**
  Play (x_t, y_t).
  Observe $\mathcal{L}_t$.
  Set $x_{t+1} = \arg\min_{x\in X}\langle \frac{\gamma_t\theta_t}{\beta_X}\nabla_x\mathcal{L}_t(x_t,y_t), x\rangle + B_{R_X}(x, x_t)$.
  Set $y_{t+1} = \arg\max_{y\in Y}\langle \frac{\gamma_t\theta_t}{\beta_Y}\nabla_y\mathcal{L}_t(x_t,y_t), y\rangle - B_{R_Y}(y, y_t)$.
end for

Theorem 2 (WOSP-Gap Bound)**.** *Let* $\Omega = \beta_X \max_{x\in X} B_{R_X}(x, x_1) + \beta_Y \max_{y\in Y} B_{R_Y}(y, y_1)$. *Setting* $\gamma_t = \sqrt{\frac{2\Omega}{G^2 T \max_{t\in[T]}\theta_t^2}}$, *the iterates* $\{(x_t, y_t)\}_{t=1}^{T}$ *generated by the Online Generalized Mirror Descent algorithm have the following WOSP-gap bound:*

$$\max_{y\in Y}\sum_{t=1}^{T}\theta_t\mathcal{L}_t(x_t, y) - \min_{x\in X}\sum_{t=1}^{T}\theta_t\mathcal{L}_t(x, y_t) \le \sqrt{2\Omega G^2 T \max_{t\in T}\theta_t^2}.$$

Therefore, for $\theta_t = \frac{1}{T}\ \forall t \in [T]$, *we have WOSP-gap* $\le \mathcal{O}(\frac{1}{\sqrt{T}})$.
Moreover, if for all $t \in [T]$, $\mathcal{L}_t(x,y) - \alpha R_X(x) + \alpha R_Y(y)$ *is convex in* x *and concave in* y, *setting* $\theta_t = \frac{2t}{T(T+1)}$ *for* $t \in [T]$, *the WOSP-gap is upper bounded by* $\frac{2G^2}{\alpha(T+1)} = \mathcal{O}(\frac{1}{T})$.

The $\mathcal{O}(\frac{1}{T})$ WOSP-gap bounds in the strongly convex case is a significant improvement over the $\mathcal{O}(\frac{\ln T}{T})$ individual regret bounds in the standard OCO literature and is a result of extra flexibility in the setup due to using non-uniform weights in the regret metric.
Next, I will talk about the Saddle Point Regularized Follow The Leader (SP-RFTL) algorithm

proposed by [1, 2]. Unlike the Online Generalized Mirror Descent algorithm, this algorithm focuses on optimizing the NE-regret through applying the seminal Regularized Follow the Leader (RFTL) of the OCO literature to the OSP problem. Note that in this algorithm, $\{(x_t, y_t)\}_{t=1}^T$ are updated jointly at each round. As $T \to \infty$, the last iterate (x_{T+1}, y_{T+1}) will converge to the Nash Equilibrium of the average game $\frac{1}{T}\sum_{t=1}^T \mathcal{L}_t$ because for $\eta = \tilde{\mathcal{O}}(\sqrt{T})$, we have:

$$x_{T+1} = \arg\min_{x\in X}\max_{y\in Y} \frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(x,y) + \underbrace{\frac{1}{\sqrt{T}}R_X(x) - \frac{1}{\sqrt{T}}R_Y(y)}_{\to 0 \text{ as } T\to\infty} = \arg\min_{x\in X}\max_{y\in Y} \frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(x,y),$$

$$y_{T+1} = \arg\max_{y\in Y}\min_{x\in X} \frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(x,y) + \underbrace{\frac{1}{\sqrt{T}}R_X(x) - \frac{1}{\sqrt{T}}R_Y(y)}_{\to 0 \text{ as } T\to\infty} = \arg\max_{y\in Y}\min_{x\in X} \frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(x,y).$$

Algorithm Saddle Point Regularized Follow The Leader (SP-RFTL)

Input: $x_1 \in X, y_1 \in Y, \eta > 0$, strongly convex functions R_X and R_Y.
Output: $\{(x_t, y_t) : 1 \le t \le T\}$.
for $t = 1$ **to** T **do**
  Play (x_t, y_t).
  Observe $\mathcal{L}_t$.
  Set $x_{t+1} = \arg\min_{x\in X}\max_{y\in Y}\sum_{s=1}^t \mathcal{L}_s(x,y) + \frac{t}{\eta}R_X(x) - \frac{t}{\eta}R_Y(y)$.
  Set $y_{t+1} = \arg\max_{y\in Y}\min_{x\in X}\sum_{s=1}^t \mathcal{L}_s(x,y) + \frac{t}{\eta}R_X(x) - \frac{t}{\eta}R_Y(y)$.
end for

Theorem 3 (NE-Regret Bound)**.** *Let G_{R_X}, G_{R_Y} be the Lipschitz constants of R_X, R_Y with respect to the norm $\|\cdot\|$. Setting $\eta = \frac{\sqrt{T}}{\ln T}$, the iterates $\{(x_t, y_t)\}_{t=1}^T$ generated by the SP-RFTL algorithm have the following NE-Regret bound:*

$$\textit{NE-Regret}(T) = |\frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(x_t, y_t) - \min_{x\in X}\max_{y\in Y}\frac{1}{T}\sum_{t=1}^T \mathcal{L}_t(x,y)| \le \mathcal{O}(\sqrt{\frac{\ln T}{T}}).$$

5 Online Matrix Games (OMG): Bandit feedback

In this section, I will focus on a special class of OSP problems, called the Online Matrix Games (OMG), where $\mathcal{L}_t(x,y) = x^T A_t y,\ A_t \in [-1,1]^{d_1\times d_2}$ for all $t \in [T]$, $X \subset \mathbb{R}^{d_1}$ and $Y \subset \mathbb{R}^{d_2}$ are the simplex. In the bandit feedback setting, both players only observe the payoff $\mathcal{L}_t(x_t, y_t)$ as the feedback at each round $t \in [T]$ and no further information about the payoff matrix A_t is observed. [1, 2] proposed the Bandit Online Matrix Games Regularized Follow the Leader (Bandit-OMG-RFTL) algorithm for this setting. The algorithm is a fairly straightforward application of the SP-RFTL algorithm using the unbiased estimators $\{\hat{A}_t\}_{t=1}^T$ of the true payoff matrices $\{A_t\}_{t=1}^T$ in the payoff functions. In online problems where the constraint sets are the simplex, the negative entropy is commonly used as the regularizer function. However, as Theorem 3 suggests, the regularizer functions need to be Lipschitz continuous to use the SP-RFTL algorithm and the negative entropy function is indeed not Lipschitz over the simplex. As a remedy, the Bandit-OMG-RFTL algorithm further restricts the two players to play over a shrunken version of the simplex defined below:

$$\Delta_\delta = \{z \in \mathbb{R}^d : \|z\|_1 = 1, z_i \ge \delta\ \forall i \in [d]\}$$

for some $0 \le \delta \le \frac{1}{d}$. The negative entropy function is $\max\{1, |\ln\delta|\}$-Lipschitz over Δ_δ with respect to the norm $\|\cdot\|_1$.

Algorithm 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 $\eta > 0$ and $0 < \delta < \min\{\frac{1}{d_1}, \frac{1}{d_2}\}$, $R_X(x) = \sum_{i=1}^{d_1} x_i \ln x_i + \ln d_1$ and $R_Y(y) = \sum_{i=1}^{d_2} y_i \ln y_i + \ln d_2$.
for $t = 1$ **to** T **do**
  Sample independently $i_t \sim x_t$ and $j_t \sim y_t$.
  Observe $[A_t]_{i_t,j_t}$.
  Set $[\hat{A}_t]_{i,j} = \frac{[A_t]_{i,j}}{[x_t]_i [y_t]_j}$ if $i = i_t, j = j_t$ and 0 otherwise.
  Set $\mathcal{L}_t(x, y) = x^T \hat{A}_t y$.
  Set $x_{t+1} = \arg\min_{x \in \Delta_{X,\delta}} \max_{y \in \Delta_{Y,\delta}} \sum_{s=1}^{t} \mathcal{L}_s(x, y) + \frac{t}{\eta} R_X(x) - \frac{t}{\eta} R_Y(y)$.
  Set $y_{t+1} = \arg\max_{y \in \Delta_{Y,\delta}} \min_{x \in \Delta_{X,\delta}} \sum_{s=1}^{t} \mathcal{L}_s(x, y) + \frac{t}{\eta} R_X(x) - \frac{t}{\eta} R_Y(y)$.
end for

Theorem 4. *Let $\{A_t\}_{t=1}^T$ be any sequence of payoff matrices chosen by an adaptive adversary. Let $\{(i_t, j_t)\}_{t=1}^T$ be the iterates generated by the Bandit-OMG-RFTL algorithm. Setting $\delta = T^{-1/6}$ and $\eta = T^{1/6}$, we have:*

$$|\mathbb{E}\big[\sum_{t=1}^{T} [A_t]_{i_t,j_t} - \min_{x \in X} \max_{y \in Y} \sum_{t=1}^{T} x^T A_t y\big]| \leq \mathcal{O}((d_1 + d_2) T^{5/6}),$$

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

6 Conclusion and future directions

In this project, I studied an extension of classical zero-sum games to the online setting where the payoff functions are evolving arbitrarily (and possibly adversarially). I went over the few papers that have considered this problem in the full information and bandit feedback settings and discussed the two main performance metrics that are commonly used in the literature along with the algorithms that aim to optimize each metric. As mentioned in Section 3.1.1, these two notions of regret are incompatible and each of the proposed algorithms in the prior works aims to obtain sub-linear bounds for one of the two metrics. In this topic, there are a number of interesting research directions to pursue in future works. Firstly, the NE-regret bound in Theorem 4 for the bandit feedback setting is sub-optimal and better bounds could be derived using more efficient estimation techniques for the payoff matrices. Moreover, the SP-RFTL algorithm involves solving a saddle point problem at each round which may not be computationally efficient and designing faster algorithms with better running times are desirable. Lastly, all the prior works have focused on the convex setting where for all $t \in [T]$, the payoff function $\mathcal{L}_t(x, y)$ is convex in x and concave in y. Extending this analysis to particular classes of non-convex problems (such as DR-submodular functions) is yet to be done.

References

[1] Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. Competing against equilibria in zero-sum games with evolving payoffs. *arXiv preprint arXiv:1907.07723*, 2019.

[2] Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. Competing against nash equilibria in adversarially changing zero-sum games. In *International Conference on Machine Learning*, pages 921–930. PMLR, 2019.

[3] Nam Ho-Nguyen and Fatma Kılınç-Karzan. Exploiting problem structure in optimization under uncertainty via online convex optimization. *Mathematical Programming*, 177(1):113–147, 2019.

[4] Adrian Rivera, He Wang, and Huan Xu. The online saddle point problem: Applications to online convex optimization with knapsacks. *arXiv preprint arXiv:1806.08301*, 2018.

[5] Abhishek Roy, Yifang Chen, Krishnakumar Balasubramanian, and Prasant Mohapatra. Online and bandit algorithms for nonstationary stochastic saddle-point optimization. *arXiv preprint arXiv:1912.01698*, 2019.

[6] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 928–935. AAAI Press, 2003.

[7] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003.

[8] Sébastien Bubeck. Introduction to online optimization. *Lecture Notes*, 2, 2011.

[9] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

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

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

[12] Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

[13] Michael Bowling and Manuela Veloso. Convergence of gradient dynamics with a variable learning rate. In *ICML*, pages 27–34, 2001.

[14] Michael Bowling. Convergence and no-regret in multiagent learning. *Advances in neural information processing systems*, 17:209–216, 2005.

[15] Vincent Conitzer and Tuomas Sandholm. 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.

[16] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1):2503–2528, 2012.

[17] Rodolphe Jenatton, Jim Huang, Dominik Csiba, and Cedric Archambeau. Online optimization and regret guarantees for non-additive long-term constraints. *arXiv preprint arXiv:1602.05394*, 2016.

[18] Alec Koppel, Felicia Y Jakubiec, and Alejandro Ribeiro. A saddle point algorithm for networked online convex optimization. *IEEE Transactions on Signal Processing*, 63(19):5149–5164, 2015.

[19] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 202–219. IEEE, 2019.