Fast and Furious Learning in Zero-Sum Games: Vanishing Regret with Non-Vanishing Step Sizes

James P. Bailey and Georgios Piliouras

Texas A&M University and Singapore University of Technology and Design

Central Question

It is well known that standard online optimization algorithms (e.g. Online Gradient Descent) can achieve sublinear $O(\sqrt{T})$ regret as long as we apply a decreasing step-size (learning rate) of $1/\sqrt{T}$ and that this bound is tight.

*Can we prove **sublinear regret** for OGD **without decreasing step-sizes** in zero-sum games?*

Significance

- Min-max optimization is a problem of interest in several communities including Optimization, Game Theory and Machine Learning.
- Gradient descent is the most well studied optimization technique and (bilinear) zero-sum games are the most standard games. Even in 2x2 games, this toy benchmark is **not well understood** and the behavior is **not simple**.
- Gradient descent in zero-sum games is actually **unstable** and **chaotic**.
- Sublinear regret implies time average convergence to exact Nash equilibria in zero-sum games.

Zero-Sum Games

A **two-player game** consists of two players $\{1, 2\}$ where each player has n_i strategies to select from. Player i can either select a pure strategy $j \in [n_i]$ or a mixed strategy $x_i \in \mathcal{X}_i = \{x_i \in \mathbb{R}^{n_i}_{\geq 0} : \sum_{j\in[n_i]} x_{ij} = 1\}$.

In **a zero-sum game**, there is a payoff matrix $A \in \mathbb{R}^{n_1 \times n_2}$ where player 1 receives utility $x_1 \cdot Ax_2$ and player 2 receives utility $-x_1 \cdot Ax_2$ resulting in the following optimization problem:

$$\max_{x_1\in\mathcal{X}_1} \min_{x_2\in\mathcal{X}_2} x_1 \cdot Ax_2 \qquad \text{(Two-Player Zero-Sum Game)}$$

The **solution** to this saddle problem is the **Nash equilibrium** x^{NE}.

$$x_1^{NE} \cdot Ax_2 \geq x_1^{NE} \cdot Ax_2^{NE}$$

independent of what strategy player 2 selects.

Full version at `arxiv:1905.04532`

Gradient-Descent-Ascent

The most common class of online learning algorithms is again the family of follow-the-regularized-leader algorithms.

$$y_1^T = y_1^0 + \sum_{t=1}^{T-1} Ax_2^t \qquad \text{(Player 1 Payoff Vector)}$$

$$y_2^T = y_2^0 - \sum_{t=1}^{T-1} A^\mathsf{T} x_1^t \qquad \text{(Player 2 Payoff Vector)}$$

$$x_i^t = \operatorname*{arg\,max}_{x_i\geq 0:\sum_{j\in[n_i]} x_{ij}=1} \left\{ y_i^t \cdot x_i - \frac{h_i(x_i)}{\eta} \right\} \qquad \text{(FTRL)}$$

$$x_i^t = \operatorname*{arg\,max}_{x_i\geq 0:\sum_{j\in[n_i]} x_{ij}=1} \left\{ y_i^t \cdot x_i - \frac{||x_i||_2^2}{2\eta} \right\} \qquad \text{(OGD)}$$

where η corresponds to the learning rate.

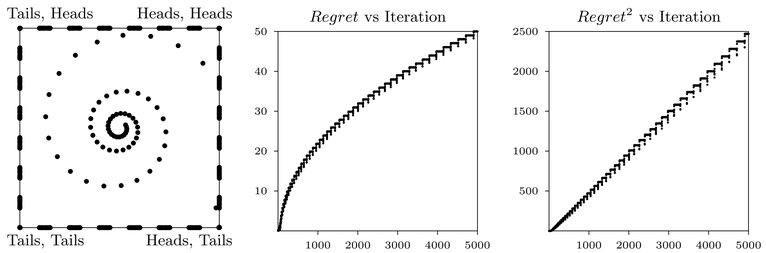
Theorem (Convergence to the Boundary)

Theorem: *Let A be a 2x2 game that has a unique fully mixed Nash equilibrium where strategies are updated according to OGD. For any non-equilibrium initial strategies and any fixed learning rate η, there exists a B such that x^t is on the boundary for all $t \geq B$.*

Theorem ($\Theta(\sqrt{T})$ Regret in 2x2 Zero-Sum Games)

Let A be a 2x2 game that has a unique fully mixed Nash equilibrium. When x^t is updated according to OGD with any fixed learning rate η, $Regret_1(T) \in O\left(\sqrt{T}\right)$.

Experiments

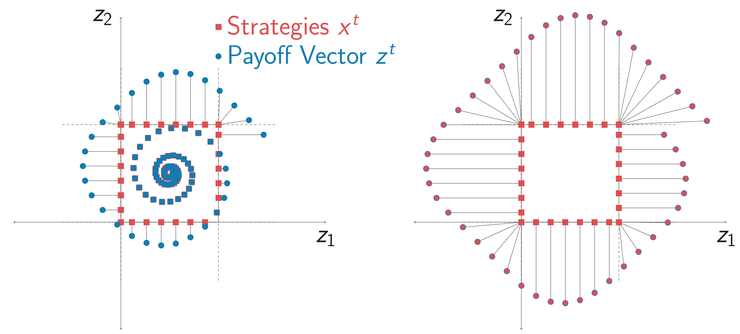


(a) Player Strategies (b) Player 1 Regret (c) Player 1 Regret Squared

Figure: 5000 Iterations of Gradient Descent on Matching Pennies with $\eta = .15$.

Understanding the Geometry of the Dynamics

The payoff vector y_i^t is a formal dual of the strategy x_i^t. We choose a dual space that will be convenient for showing our results in 2x2 zero-sum games.



(a) Iterations 1-95 (b) Iterations 95-140

Figure: Strategies and Transformed Payoff Vectors Rotating Clockwise and Outwards in Matching Pennies with $\eta = .15$ and $(y_{11}^0, y_{11}^0) = (.2, -.3)$.

Key Proof Idea

The **sum of the convex conjugates of the regularizers** can be thought as a **non-decreasing system "energy"**. We keep track of its increase by partitioning the space in regions.

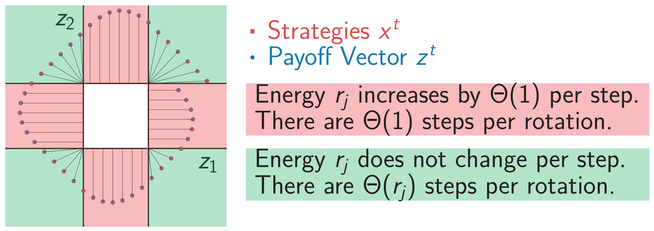


Figure: Partitioning of Payoff Vectors

As a result, strategies are almost always in the corners.

$$Regret_1(T) \leq O(1) + \sum_{t=0}^{T} (x_1^{t+1} - x_1^t) \cdot Ax_2^t \qquad (1)$$

Since the strategies rarely change, the regret rarely grows resulting in $O(\sqrt{T})$ regret.

The proof ideas extend to higher dimensions and to other variants of FTRL. Moreover, we provide experimental evidence that sublinear regret extends to these settings.