15-888 Computational Game Solving (Fall 2025) Monday, September 22th, 2025

Lecture 8
Faster No-Regret Learning Dynamics and Last-Iterate Convergence

Ioannis Anagnostides

A basic premise in the online learning framework is that the sequence of utilities given as input to the learner is produced adversarially, so as to maximize the regret of the learner. This worst-case perspective can be overly pessimistic, particularly for the main application of regret minimization we have seen in the course—self-play. Indeed, when we design algorithms in self-play, we have considerable control concerning the underlying sequence of utilities. Can we overcome the $\sqrt{T}$ barrier ingrained in the adversarial setting? Answering this question is the main subject of this lecture.

In more detail, Section 1 introduces the framework of *optimistic* no-regret learning, which is predicated on using *predictions*. All online algorithms we have covered so far have an optimistic counterpart. We will see how to obtain near-optimal rates first in (two-player) zero-sum games (Section 1.1) and then in general-sum multi-player games (Section 1.2). Now, regret-based guarantees translate to some form of time-average convergence. But what can be said about the *last iterate* of those algorithms? This is addressed in Section 2. While traditional algorithms, such as `MWU`, fail miserably to converge in a last-iterate sense, their optimistic counterparts *do* converge at least in zero-sum games. In other words, there is a strong connection between last-iterate convergence and near-optimal regret.

1 Optimistic no-regret learning

As we saw in an earlier lecture, a learner can always guarantee regret bounded as $\sqrt{T}$ in terms of the time horizon T. How can we improve that when the environment is more benign? The *optimistic* (or *predictive)* online learning framework provides an answer to this question. The key idea is to incorporate some form of *prediction* regarding the next utility.

Necessity of optimism Before we formally introduce the optimistic framework, it's worth first highlighting the intrinsic limitations of non-optimistic no-regret algorithms. A priori, one possibility would be that one can come up with a better analysis to show that some of the algorithms that we have already seen, such as multiplicative weights update (`MWU`), will have smaller regret in self-play; after all, the usual $\Omega(\sqrt{T})$ lower bound is very much not compatible with the adversary running `MWU`. Yet, this turns out to be impossible.

Theorem 1.1 (Chen and Peng, 2020)**.** *For any learning rate, when both players in a two-player game employ* `MWU`*, at least one player will have* $\Omega(\sqrt{T})$ *regret.*

The main takeaway is that we will need new algorithmic ideas to overcome the $\sqrt{T}$ barrier; improving our analysis will not be enough. This is where optimism comes into play [Chiang et al.,

2012, Rakhlin and Sridharan, 2013]. Optimistic algorithms maintain, for each round $t \in [T]$, a prediction vector $\boldsymbol{m}^{(t)}$ that is being updated dynamically; the goal of the prediction is to match as closely as possible the next utility $\boldsymbol{u}^{(t)}$.

All online algorithms we have seen have an optimistic counterpart. First, the optimistic version of FTRL (henceforth, OFTRL) is given by

$$\boldsymbol{x}^{(t)} = \operatorname*{argmax}_{\boldsymbol{x} \in \mathcal{X}} \left\{ \left\langle \boldsymbol{x}, \boldsymbol{m}^{(t)} + \sum_{\tau=1}^{t-1} \boldsymbol{u}^{(\tau)} \right\rangle - \frac{1}{\eta} \mathcal{R}(\boldsymbol{x}) \right\}. \tag{1}$$

We recall that $\eta > 0$ is the learning rate—a parameter that can be tuned appropriately; $\mathcal{R}$ is a strictly convex regularizer, such as the squared Euclidean norm or (negative) entropy; and $\mathcal{X}$ is a convex and compact set of strategies, such as the sequence-form polytope. The only difference of (1) relative to FTRL is that it incorporates a prediction vector $\boldsymbol{m}^{(t)}$. In particular, if one sets $\boldsymbol{m}^{(t)} = \mathbf{0}$ for all t, one reverts to FTRL. OFTRL was introduced by Syrgkanis et al. [2015].

In a similar vein, the optimistic version of MD [Chiang et al., 2012, Rakhlin and Sridharan, 2013] (abbreviated as OMD, not to be confused with online mirror descent) is defined as

$$\boldsymbol{x}^{(t)} \coloneqq \operatorname*{argmax}_{\boldsymbol{x} \in \mathcal{X}} \left\{ \langle \boldsymbol{x}, \boldsymbol{m}^{(t)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\boldsymbol{x}, \hat{\boldsymbol{x}}^{(t-1)}) \right\},$$

where

$$\hat{\boldsymbol{x}}^{(t)} \coloneqq \operatorname*{argmax}_{\hat{\boldsymbol{x}} \in \mathcal{X}} \left\{ \langle \hat{\boldsymbol{x}}, \boldsymbol{u}^{(t)} \rangle - \frac{1}{\eta} \mathcal{B}_{\mathcal{R}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}^{(t-1)}) \right\}.$$

We recall that $\mathcal{B}_{\mathcal{R}}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{x}') - \langle \nabla \mathcal{R}(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle$ is the Bregman divergence induced by $\mathcal{R}$. Unlike OFTRL, OMD relies on a secondary, auxiliary sequence of strategies $(\hat{\boldsymbol{x}}^{(t)})_{t=1}^{T}$.

Now, there are many different ways to set the prediction vector, all based on taking some convex combination of previously observed utilities; this could be based on taking the uniform average over a fixed window H, or geometrically discounting past utilities. By far the most common prediction mechanism is a simple one-step recency bias, $\boldsymbol{m}^{(t)} \coloneqq \boldsymbol{u}^{(t-1)}$. In the sequel, we always use this simple prediction unless explicitly stated otherwise.

The analysis of optimistic no-regret learning revolves around a key property referred to as *regret bounded by variation in utilities (RVU)*.

Definition 1.2 (RVU; Syrgkanis et al., 2015)**.** An online algorithm satisfies the *RVU* property if there are constants $\alpha, \beta, \gamma > 0$ such that its regret can be bounded as

$$\mathrm{Reg}^{(T)} \leq \alpha + \beta \sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_*^2 - \gamma \sum_{t=1}^{T} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|^2. \tag{2}$$

Compared to the usual regret bound of algorithms such as FTRL or MD, there are two notable differences in (2). The first one is that the regret is bounded by the *misprediction error*, measured by $\sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_*^2$. The crux of the entire analysis is to show that, in self-play, the misprediction error will be small. The second and more enigmatic difference in (2) is the negative term, indicating that *large variation* in the player's strategies results in smaller regret. Both OFTRL and OMD enjoy the RVU property with the following set of parameters.

Proposition 1.3 (Syrgkanis et al., 2015). *Both* OFTRL *and* OMD *have the RVU property with* $\alpha \propto 1/\eta$, $\beta = \eta$, *and* $\gamma \propto 1/\eta$.

As a result, if one ignores the negative term in the RVU property for now, the optimal regret bound as a function of η is proportional $\sqrt{\sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_*^2}$. This is always bounded by $\sqrt{T}$, even if the predictions are grossly inaccurate, but it can be much smaller when the misprediction error is small.

Predictive regret matching One can also incorporate predictions in the regret matching framework [Farina et al., 2021]. The resulting algorithms are given in Algorithms 1 and 2. The overarching idea is again the same. One uses a prediction $\boldsymbol{m}^{(t)}$ so as to guess the next regret vector. Farina et al. [2021] have shown that both algorithms guarantee regret bounded as $\sum_{t=1}^{T} \|\boldsymbol{u}^{(t)} - \boldsymbol{m}^{(t)}\|_*^2$, just like OFTRL and OMD; on the flip side, neither PRM nor PRM$^+$ satisfies the RVU property. Yet, as we shall see in more detail in the next lecture, PRM and PRM$^+$ are some of the most effective algorithms for solving zero-sum games.

Algorithm 1: Predictive RM (PRM)

Initialize cumulative regrets $\boldsymbol{r}^{(0)} \coloneqq \mathbf{0}$;
for $t = 1, \dots, T$ **do**
  Define $\boldsymbol{\theta}^{(t)} \coloneqq [\boldsymbol{r}^{(t-1)} + \boldsymbol{m}^{(t)} - \langle \boldsymbol{m}^{(t)}, \boldsymbol{x}^{(t-1)} \rangle \mathbf{1}]^+$;
  if $\boldsymbol{\theta}^{(t)} = \mathbf{0}$ **then**
    Let $\boldsymbol{x}^{(t)} \in \Delta(\mathcal{A})$ be arbitrary
  else
    Compute $\boldsymbol{x}^{(t)} \coloneqq \boldsymbol{\theta}^{(t)} / \|\boldsymbol{\theta}^{(t)}\|_1$;
  Output strategy $\boldsymbol{x}^{(t)} \in \Delta(\mathcal{A})$;
  Observe utility $\boldsymbol{u}^{(t)} \in [0,1]^{\mathcal{A}}$;
  $\boldsymbol{r}^{(t)} \coloneqq \boldsymbol{r}^{(t-1)} + \boldsymbol{u}^{(t)} - \langle \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle \mathbf{1}$;

Algorithm 2: Predictive RM$^+$ (RM$^+$)

Initialize cumulative regrets $\boldsymbol{r}^{(0)} \coloneqq \mathbf{0}$;
for $t = 1, \dots, T$ **do**
  Define $\boldsymbol{\theta}^{(t)} \coloneqq [\boldsymbol{r}^{(t-1)} + \boldsymbol{m}^{(t)} - \langle \boldsymbol{m}^{(t)}, \boldsymbol{x}^{(t-1)} \rangle \mathbf{1}]^+$;
  if $\boldsymbol{\theta}^{(t)} = \mathbf{0}$ **then**
    Let $\boldsymbol{x}^{(t)} \in \Delta(\mathcal{A})$ be arbitrary
  else
    Compute $\boldsymbol{x}^{(t)} \coloneqq \boldsymbol{\theta}^{(t)} / \|\boldsymbol{\theta}^{(t)}\|_1$;
  Output strategy $\boldsymbol{x}^{(t)} \in \Delta(\mathcal{A})$;
  Observe utility $\boldsymbol{u}^{(t)} \in [0,1]^{\mathcal{A}}$;
  $\boldsymbol{r}^{(t)} \coloneqq [\boldsymbol{r}^{(t-1)} + \boldsymbol{u}^{(t)} - \langle \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle \mathbf{1}]^+$;

1.1 Near-optimal regret in zero-sum games

We will now see how to leverage the RVU property to obtain near-optimal regret bounds, focusing first on (two-player) zero-sum games. The first point to make here is that it's impossible to guarantee regret smaller than a constant. This can be seen even in a single-player (static) setting: the learner, who has no prior information, will likely fail to identify the optimal strategy in the first round, which means that its regret will be $\Omega(1)$ even if the learner plays optimally from the second round onward. As a result, our goal will be to guarantee bounded regret.

To begin with, we point out two key properties. The first is that two consecutive strategies of OFTRL or OMD are close to each other, which is a consequence of regularization.

Lemma 1.4 (Stability). *For any round* t, OFTRL *and* OMD *guarantee* $\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\| \le O(\eta)$.

In fact, this stability property is essential to prove the $\sqrt{T}$ bound in the adversarial setting. It is *not* satisfied for algorithms such as fictitious play or best-response dynamics, or even regret matching. The issue with regret matching occurs when all regrets are either negative or close to zero; then, even a tiny utility vector can result in entirely different strategies. This is the basic reason why the upcoming analysis—or any analysis for that matter [Farina et al., 2023]—does not work for PRM and PRM$^+$.

The second key property is that the sequence of utilities observed by each player is Lipschitz continuous in terms of the strategies employed by the other players; the lemma as stated below is specifically for normal-form games, but similar bounds hold more generally without any qualitative differences.

Lemma 1.5. *For any player $i \in [n]$, $\|\boldsymbol{u}_i^{(t)} - \boldsymbol{u}_i^{(t-1)}\|_\infty \le \sum_{i' \neq i} \|\boldsymbol{x}_{i'}^{(t)} - \boldsymbol{x}_{i'}^{(t-1)}\|_1$, where $\boldsymbol{u}_i^{(t)} = \boldsymbol{u}_i(\boldsymbol{x}_{-i}^{(t)})$.*

We recall from a previous lecture that $\boldsymbol{u}_i(\boldsymbol{x}_{-i}) = (\mathbb{E}_{a_{-i} \sim \boldsymbol{x}_{-i}} u_i(a_i, a_{-i}))_{a_i \in \mathcal{A}_i}$ is the utility gradient of player i. To obtain Lemma 1.5, one can observe that

$$\begin{aligned}\left| \sum_{a_{-i}} \prod_{i' \neq i} \boldsymbol{x}_{i'}^{(t)}[a_{i'}] u_i(\cdot, a_{-i}) - \sum_{a_{-i}} \prod_{i' \neq i} \boldsymbol{x}_{i'}^{(t-1)}[a_{i'}] u_i(\cdot, a_{-i}) \right| &\le \sum_{a_{-i}} \left| \prod_{i' \neq i} \boldsymbol{x}_{i'}^{(t)}[a_{i'}] - \prod_{i' \neq i} \boldsymbol{x}_{i'}^{(t-1)}[a_{i'}] \right| \\ &\le \sum_{i' \neq i} \|\boldsymbol{x}_{i'}^{(t)} - \boldsymbol{x}_{i'}^{(t-1)}\|_1,\end{aligned}$$

where the first inequality uses the triangle inequality together with the assumption that $|u_i(\cdot)| \le 1$, and the second inequality follows because the total variation distance between two product distributions is bounded in terms of the sum of the total variation distances of the marginals.

Let's think of what Lemmas 1.4 and 1.5 imply together with the RVU property. We know that OFTRL and OMD perform well when the misprediction error is small. By Lemma 1.5, in the self-play setting, the misprediction error is bounded by the variation of the other players' strategies. But Lemma 1.4 tells us that if *all* players follow a stable algorithm, such as OFTRL or OMD, the misprediction error will be smaller than what could be obtained in the adversarial setting.

Theorem 1.6 (Syrgkanis et al., 2015)**.** *If all players employ* OFTRL *or* OMD *in a multi-player setting with learning rate $\eta = T^{-1/4}$, the regret of each player is bounded as $T^{1/4}$.*

Proof. Combining Proposition 1.3 with Lemmas 1.4 and 1.5, we have that the regret of each player $i \in [n]$ as a function of η and T is bounded as

$$\mathrm{Reg}_i^{(T)} \le O\left(\frac{1}{\eta}\right) + \eta \sum_{t=1}^{T} \|\boldsymbol{u}_i^{(t)} - \boldsymbol{u}_i^{(t-1)}\|_*^2 \le O\left(\frac{1}{\eta}\right) + O(\eta^3 T). \tag{3}$$

Optimizing over the learning rate η gives the claim. □

The regret bound in (3) is to be compared with the bound $O(1/\eta) + O(\eta T)$ obtained in the adversarial setting.

While simply using the stability property already gets a significant improvement of $T^{1/4}$, we are still far from our goal of getting constant regret. To do so, we will need to find a way to make use of the negative term in the RVU bound.

Let's assume that all players employ an algorithm that satisfies the RVU property with respect to the parameters (α, β, γ). A key observation is that, by Lemma 1.5,

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{Reg}_i^{(T)} &\le \alpha n + (n-1)\beta \sum_{i=1}^{n} \sum_{i' \neq i} \sum_{t=1}^{T} \|\boldsymbol{x}_{i'}^{(t)} - \boldsymbol{x}_{i'}^{(t)}\|_1^2 - \gamma \sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \\
&\le \alpha n + (n-1)^2 \beta \sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t)}\|_1^2 - \gamma \sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \\
&\le \alpha n + \sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \left((n-1)^2 \beta - \gamma\right) \\
&\le \alpha n - \frac{\gamma}{2} \sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2, \qquad (4)
\end{aligned}$$

so long as $\gamma \ge 2(n-1)^2\beta$. Further, by Proposition 1.3, both `OFTRL` and `MD` satisfy the RVU property with $\gamma \propto 1/\eta$ and $\beta = \eta$; so, taking $\eta \propto 1/n$, we can guarantee $\sum_{i=1}^{n} \mathrm{Reg}_i^{(T)} = O(1)$. However, our goal is to bound the *maximum* of the players' regrets, not the sum. The issue is that a player can have $\Omega(T)$ regret even though the sum of the players' regrets is bounded; this can happen if some other player has regret $-\Theta(T)$, which ends up canceling the large regret. To address this discrepancy, for now, we focus on the class of games in which the *sum of the players' regrets is nonnegative*, $\sum_{i=1}^{n} \mathrm{Reg}_i^{(T)} \ge 0$.

Games with nonnegative sum of regrets The canonical example are two-player zero-sum games. As we saw in an earlier lecture, for any sequence of strategies $(\boldsymbol{x}^{(t)})_{t=1}^{T}$ and $(\boldsymbol{y}^{(t)})_{t=1}^{T}$, the sum of the two players' regrets is equal to the *duality gap*, which is in turn nonnegative:

$$\begin{aligned}
\mathrm{Reg}_1^{(T)} + \mathrm{Reg}_2^{(T)} &= \max_{\boldsymbol{x}' \in \mathcal{X}} \sum_{t=1}^{T} \langle \boldsymbol{x}' - \boldsymbol{x}^{(t)}, -\mathbf{A}\boldsymbol{y}^{(t)} \rangle + \max_{\boldsymbol{y}' \in \mathcal{Y}} \sum_{t=1}^{T} \langle \boldsymbol{y}' - \boldsymbol{y}^{(t)}, \mathbf{A}^\top \boldsymbol{x}^{(t)} \rangle \\
&= T \left(\max_{\boldsymbol{y}' \in \mathcal{Y}} \langle \boldsymbol{y}', \mathbf{A}^\top \bar{\boldsymbol{x}}^{(T)} \rangle - \min_{\boldsymbol{x}' \in \mathcal{X}} \langle \boldsymbol{x}', \mathbf{A}\bar{\boldsymbol{y}}^{(T)} \rangle \right) \ge 0.
\end{aligned}$$

There are also interesting classes of multi-player games that have this property, such as *polymatrix zero-sum* games [Cai et al., 2016]; this is a generalization of two-player zero-sum games. It is a based on an underlying graph. Each player is uniquely associated to a node in the graph. Each edge corresponds to a (two-player) zero-sum game played between the incident players. The utility of a player is given by the sum of the utilities with respect to the corresponding games.

Under that assumption, rearranging from (4) implies the following important consequence.

Theorem 1.7 (Anagnostides et al., 2022a)**.** *If all players employ* OFTRL *or* OMD *with a sufficiently small learning rate in a game with nonnegative sum of regrets,*

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \leq O(1). \tag{5}$$

The left-hand side of (5) is called *second-order path length*. Theorem 1.7 implies that OFTRL or OMD guarantee a bounded second-order path length. (We will see later what that property tells us about last-iterate convergence.) In turn, this implies that the misprediction error of each player will be bounded, so all players will have bounded regret.

Corollary 1.8. *If all players employ* OFTRL *or* OMD *with a sufficiently small learning rate in a game with nonnegative sum of regrets, each player has bounded regret.*

1.2 Near-optimal regret in general-sum games

The main assumption we made in Corollary 1.8 is that the sum of the players' regrets is nonnegative, which is a severe restriction on the class of games. We will now see how to strengthen the previous approach. The basic idea is to consider a stronger notion of regret, one that happens to be always *nonnegative*. One such notion of regret is *swap regret*, denoted by $\mathsf{SwapReg}^{(T)}$, introduced in an earlier lecture.

Observation 1.9. *For any sequence of utilities* $(\boldsymbol{u}^{(t)})_{t=1}^T$ *and strategies* $(\boldsymbol{x}^{(t)})_{t=1}^T$*, the swap regret*

$$\mathsf{SwapReg}^{(T)} = \max_{\phi \in \Phi_{\text{swap}}} \sum_{t=1}^{T} \langle \phi(\boldsymbol{x}^{(t)}) - \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle$$

is nonegative.

Above, Φ_{swap} is the set of all swap deviations. The fact that $\mathsf{SwapReg}^{(T)} \geq 0$ follows simply because Φ_{swap} contains the identity mapping. In particular, instead of swap regret, one could instead work with $\max\{0, \mathsf{Reg}^{(T)}\}$, which can be thought of as a notion of Φ-regret in which Φ contains all constant transformations together with the identity. In what follows, we will perform the analysis for the stronger notion of swap regret.

By considering the players' swap regrets, we know that their sum will be nonnegative without restricting the class of games. The crux lies in showing that there is an algorithm that guarantees an RVU bound with respect to swap regret.

To do so, we recall the algorithm of Blum-Mansour for minimizing swap regret (Algorithm 3). We know from that reduction that $\mathsf{SwapReg}^{(T)} = \sum_{a \in \mathcal{A}}^{(T)} \mathsf{Reg}_a^{(T)}$, where each $\mathsf{Reg}_a^{(T)}$ denotes the external regret of $\mathfrak{R}_a$, which operates over the ath column of the stochastic matrix. By instantiating each $\mathfrak{R}_a$ with OFTRL or OMD, it follows from Proposition 1.3 that the swap regret can be bounded as

$$\mathsf{SwapReg}^{(T)} \leq O\left(\frac{1}{\eta}\right) + \sum_{a \in \mathcal{A}} \eta \sum_{t=1}^{T} \|\boldsymbol{u}_a^{(t)} - \boldsymbol{u}_a^{(t-1)}\|_*^2 - \Omega\left(\frac{1}{\eta}\right) \sum_{a \in \mathcal{A}} \sum_{t=1}^{T} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\|^2$$

To conclude the proof, it suffices to show that

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1 \le C \sum_{a \in \mathcal{A}} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\| \tag{6}$$

for some parameter C. That is, the fixed point $\boldsymbol{x}^{(t)}$ of the Markov chain has to evolve smoothly as a function of the transition probabilities. However, this is not the case in general. Let's consider the following two stochastic matrices in the regime where $\epsilon \ll 1$.

$$\mathbf{M} = \begin{bmatrix} 1-\epsilon & 2\epsilon \\ \epsilon & 1-2\epsilon \end{bmatrix} \text{ and } \mathbf{M}' = \begin{bmatrix} 1-2\epsilon & \epsilon \\ 2\epsilon & 1-\epsilon \end{bmatrix}, \tag{7}$$

where we think of $\mathbf{M}'$ as being obtained from $\mathbf{M}$ after an update. Even though $\mathbf{M} \approx \mathbf{M}'$, the unique fixed point of $\mathbf{M}$ is $(2/3, 1/3)$ while the unique fixed point of $\mathbf{M}'$ is $(1/3, 2/3)$. So, the fixed points of two Markov chains whose transition probabilities are close to each other can still be far apart, in clear violation of (6).

There are two key ideas needed to sidestep this obstacle. The first is that stability of fixed points is guaranteed when one can ensure a stronger notion of stability for the stochastic matrix $\mathbf{M}$, namely, *multiplicative stability*. Multiplicative stability for a vector $\boldsymbol{x}_a^{(t)}$ means that

$$\max_{a' \in \mathcal{A}} \max \left\{ 1 - \frac{\boldsymbol{x}_a^{(t)}[a']}{\boldsymbol{x}_a^{(t-1)}[a']}, 1 - \frac{\boldsymbol{x}_a^{(t-1)}[a']}{\boldsymbol{x}_a^{(t)}[a']} \right\} \le O(\eta).$$

The example that we saw in (7) does not have this property: a coordinate changing from ϵ to 2ϵ violates multiplicative stability, *no matter how small* ϵ is. Most of the algorithms we have seen do not guarantee this stronger notion of stability, but it's not hard to see that MWU does. It turns out that if the transition probabilities are multiplicatively close, their fixed points will also be close. Chen and Peng [2020] made this observation to show that the Blum-Mansour algorithm in conjunction with MWU guarantee stability of the fixed points, so one can get an improved bound of $T^{1/4}$ (as in Theorem 1.6). The proof is based on the *Markov chain tree theorem*, which provides a combinatorial, closed-form solution for the stationary distribution of an ergodic Markov chain. In particular, we denote by $\mathbb{T}_a$ the set of all *directed trees* rooted at $a \in \mathcal{A}$ (the precise definition is not important for our purposes here).

Fact 1.10. *The stationary distribution $\boldsymbol{x} \in \Delta(\mathcal{A})$ of an ergodic Markov chain described through a column-stochastic matrix $\mathbf{M}$ can be expressed as*

$$\boldsymbol{x}[a] = \frac{\sum_{\mathcal{T} \in \mathbb{T}_a} \prod_{(u,v) \in E(\mathcal{T})} \mathbf{M}[v,u]}{\sum_{a' \in \mathcal{A}} \sum_{\mathcal{T} \in \mathbb{T}_{a'}} \prod_{(u,v) \in E(\mathcal{T})} \mathbf{M}[v,u]}.$$

Even so, it's still unclear whether one can get an RVU bound for swap regret through MWU. Instead, the key idea is to use a different regularizer, namely, the *logarithmic regularizer*

$$\mathcal{R} : \boldsymbol{x} \mapsto -\sum_{a \in \mathcal{A}} \log \boldsymbol{x}[a].$$

The key property of this regularizer relates to an important aspect of the regret bound of FTRL or MD we have neglected so far: the primal norm $\|\cdot\|$ under which $\mathcal{R}$ is strongly convex. It turns out that the analysis carries over even when one uses a certain non-static norm induced by the Hessian of $\mathcal{R}$, which is called *local norm*. In the case of the logarithmic regularizer, the induced local norm is

$$\|\boldsymbol{x}\|_{\boldsymbol{x}'} = \sqrt{\sum_{a\in\mathcal{A}} \left(\frac{\boldsymbol{x}[a]}{\boldsymbol{x}'[a]}\right)^2}$$

This precisely measures some form of *multiplicative* deviation. Indeed, if we use the Markov chain tree theorem, we can prove that

$$\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t-1)}\|_1 \le C \sum_{a\in\mathcal{A}} \|\boldsymbol{x}_a^{(t)} - \boldsymbol{x}_a^{(t-1)}\|_{\boldsymbol{x}_a^{(t-1)}}.$$

We thus arrive at the following RVU bound for swap regret.

Theorem 1.11 (Anagnostides et al., 2022b)**.** *There is an algorithm that guarantees the RVU property with respect to swap regret with parameters $\alpha \propto \log T/\eta$, $\beta \propto \eta$, and $\gamma \propto 1/\eta$.*

The reason we get a $\log T$ factor in the RVU bound is that the logarithmic regularizer has an *unbounded range*. One can handle that issue by considering in the definition of the regret only comparators away from the relative boundary of the simplex.

Corollary 1.12. *There is an algorithm that, when employed by all players, guarantees*

$$\sum_{i=1}^{n}\sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \le O(\log T)$$

in any general-sum multi-player game. In particular, each player will have $O(\log T)$ swap regret.

It should be noted that the first near-optimal bound on the external regret was shown fairly recently by Daskalakis et al. [2021] by analyzing the optimistic version of MWU; their analysis is quite involved, and will not be covered here.

Finally, a natural question to consider here is what happens if one or more players start deviating from the proposed protocol. Can we still guarantee the $\sqrt{T}$ regret that is possible in the adversarial setting? This is indeed possible. The basic idea is to have the player keep track of the second-order variation in the utilities,

$$\sum_{\tau=1}^{t} \|\boldsymbol{u}^{(\tau)} - \boldsymbol{u}^{(\tau-1)}\|_*^2.$$

So long as all players follow the prescribed learning algorithm, we know that $\sum_{\tau=1}^{t} \|\boldsymbol{u}^{(\tau)} - \boldsymbol{u}^{(\tau-1)}\|_*^2 = O(\log t)$. So, if the player in some round t detects that $\sum_{\tau=1}^{t} \|\boldsymbol{u}^{(\tau)} - \boldsymbol{u}^{(\tau-1)}\|_*^2 \ge \omega(\log t)$, it can revert to an algorithm tuned to work in the adversarial setting.

Algorithm 3: Blum-Mansour algorithm for minimizing swap regret

Input: A regret minimizer $\mathfrak{R}_a$ for each action $a \in \mathcal{A}$
NextStrategy():
  for *each action* $a \in \mathcal{A}$ **do**
    $\Delta(\mathcal{A}) \ni \boldsymbol{x}_a^{(t)} \coloneqq \mathfrak{R}_a$.NextStrategy();
  Set $\mathbf{M}^{(t)} \coloneqq [(\boldsymbol{x}_a^{(t)})_{a \in \mathcal{A}}]$;
  return $\Delta(\mathcal{A}) \ni \boldsymbol{x}^{(t)} = \mathbf{M}^{(t)} \boldsymbol{x}^{(t)}$;
ObserveUtility($\boldsymbol{u}^{(t)} \in \mathbb{R}^{\mathcal{A}}$):
  for *each action* $a \in \mathcal{A}$ **do**
    Set $\boldsymbol{u}_a^{(t)} \coloneqq \boldsymbol{x}^{(t)}[a] \boldsymbol{u}^{(t)}$;
    $\mathfrak{R}_a$.ObserveUtility($\boldsymbol{u}_a^{(t)}$);

2 Last-iterate convergence

All the guarantees we have seen so far based on regret minimization apply after we perform some form of averaging. In zero-sum games, the average strategies converge to the set of minimax equilibria, while more generally, it is the average correlated distribution of play that converges to the set of (coarse) correlated equilibria. This brings the question: what can be said about the last-iterate of those algorithms?

It turns out that traditional no-regret algorithms, such as MWU or online gradient descent, fail to converge in iterates to the set of equilibria. Indeed, game dynamics are known to exhibit remarkably complex behavior even in zero-sum games (for example, see Andrade et al., 2021). Yet, it turns out that optimism remedies this issue. So, while we introduced optimism as a way to obtain improved regret bounds, it incidentally also leads to last-iterate convergence in some classes of games.

Convergence in games with nonnegative sum of regrets Let's focus on the class of games with nonnegative sum of regrets. We have already shown that OMD and OFTRL guarantee

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \leq O(1).$$

This means that the strategies of each player are varying very slowly over time, perhaps suggesting some form of convergence. It turns out that for some algorithms in the OMD family—such as optimistic gradient descent, small variation does imply that the player is already best responding.

Lemma 2.1. *Suppose that each player employs* OMD *with a smooth regularizer.*[1] *Then, if* $\|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\| \leq \epsilon$ *for all players* $i \in [n]$, $\boldsymbol{x}^{(t)}$ *is an* $O(\epsilon)$*-Nash equilibrium.*

[1] That is, $\|\nabla \mathcal{R}(\boldsymbol{x}) - \nabla \mathcal{R}(\boldsymbol{x}')\|_* \leq G\|\boldsymbol{x} - \boldsymbol{x}'\|$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. This notion of smoothness is not to be confused with smoothness per Definition 2.3.

We caution that this lemma does not hold for all regularizers; a non-example is MWU. If we combine Lemma 2.1 with Theorem 1.7, we find that the individual iterates converge at a rate of $T^{-1/2}$ to approximate Nash equilibria in any game with nonnegative sum of regrets.

Corollary 2.2. *Suppose that each player employs* OMD *with a smooth regularizer in a game with nonnegative sum of regrets. For any $\epsilon > 0$, most of the strategies are ϵ-Nash equilibria after $T = O(1/\epsilon^2)$ rounds.*[2]

In light of this fact, it is perhaps not surprising that OMD has bounded regret in, for example, zero-sum games. More surprising is the behavior of the learning dynamics corresponding to Corollary 1.12 in general-sum games, illustrated in Figure 1. We see that even though players fail to converge to approximate Nash equilibria, their strategies are moving with an arbitrarily slow pace; so this algorithm does not meet the requirement to apply Lemma 2.1—the regularizer is not smooth. But the algorithm is still able to guarantee $O(\log T)$ regret for each player. As before, the misprediction error will be small even though players never actually converged to approximate Nash equilibria.

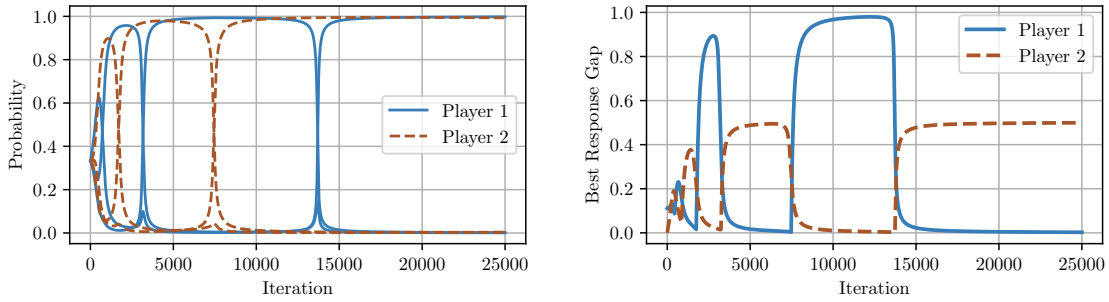

Figure 1: Taken from Anagnostides et al. [2022b].

Smooth games Finally, it's worth extending our scope beyond games with nonnegative sum of regrets. In particular, we consider the class of *smooth games*, introduced by Roughgarden [2015]. The key motivation for this definition is centered on equilibrium selection. As we have seen, a game can have multiple equilibria, and some are more desirable than others. A common metric to evaluate the quality of equilibria is the *social welfare*, $\mathrm{SW}(\boldsymbol{x}) \coloneqq \sum_{i=1}^{n} u_i(\boldsymbol{x})$. The question that arises then is how to guarantee that no-regret learning leads to outcomes whose social welfare is not too far from optimal, denoted by OPT. The following concept provides an elegant answer.

Definition 2.3 (Smooth games)**.** A game is (λ, μ)*-smooth* with respect to a welfare-optimal strategy profile $(\boldsymbol{x}_1', \dots, \boldsymbol{x}_n')$ if

$$\sum_{i=1}^{n} u_i(\boldsymbol{x}_i', \boldsymbol{x}_{-i}) \geq \lambda \mathrm{OPT} - \mu \sum_{t=1}^{n} u_i(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \quad \forall (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n).$$

[2]Proving last-iterate convergence is considerably harder [Cai et al., 2022, Gorbunov et al., 2022].

In words, a game is smooth if by having each player play its component of a welfare-optimal strategy $\boldsymbol{x}'$, players collectively guarantee some fraction of the optimal social welfare *no matter what the other players are doing*. While Definition 2.3 may seem somewhat artificial, Roughgarden [2015] showed that many interesting classes of games adhere to it. It turns out that there is a close connection between smoothness and regret minimization, formalized below.

Proposition 2.4. *Suppose that each player $i \in [n]$ has regret $\mathsf{Reg}_i^{(T)}$ in a (λ, μ)-smooth game. Then*

$$\frac{1}{T}\sum_{t=1}^{T} \mathsf{SW}(\boldsymbol{x}_1^{(t)}, \dots, \boldsymbol{x}_n^{(t)}) \geq \frac{\lambda}{1+\mu}\mathsf{OPT} - \frac{1}{1+\mu}\frac{1}{T}\sum_{i=1}^{n} \mathsf{Reg}_i^{(T)}.$$

So, in a (λ, μ)-smooth game, no-regret dynamics always secure at least a $\lambda/1+\mu =: \rho$ fraction of the optimal welfare; ρ is referred to as the *robust price of anarchy*. In particular, the rate of convergence is driven by the *sum* of the players' regrets.

Let's now think of what this means when players employ `OMD` with a smooth regularizer (per Lemma 2.1). There are two possibilities. If the sum of the players' regrets is nonnegative, we know that the players converge to a Nash equilibrium (Corollary 2.2). Otherwise, the sum of the regrets must be negative. But what this means in light of Proposition 2.4 is that not only are we matching a $\lambda/1+\mu$ of the optimal welfare, but we are in fact strictly *outperforming* it. In other words, when `OMD` fails to converge, its welfare is higher than predicted by the smoothness theory.

Theorem 2.5 (Anagnostides et al., 2022a)**.** *If all players employ* `OMD` *with a smooth regularizer for $T = O(1/\epsilon^2)$ rounds, then*

- *either the players have converged to an ϵ-Nash equilibrium, or*
- *the average social welfare is at least $\lambda/(1+\mu) + \Omega(\epsilon^2)$.*

Interestingly, the farther away from Nash equilibria the players are, the larger the improvement in terms of social welfare.

References

Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. In *Neural Information Processing Systems*, 2020.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory (COLT)*, 2012.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory (COLT)*, 2013.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Neural Information Processing Systems*, 2015.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *Conference on Artificial Intelligence (AAAI)*, 2021.

Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, and Haipeng Luo. Regret matching+: (in)stability and fast convergence in games. In *Neural Information Processing Systems*, 2023.

Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos H. Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 2016.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning (ICML)*, 2022a.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with *O(log T)* swap regret in multiplayer games. In *Neural Information Processing Systems*, 2022b.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. In *Neural Information Processing Systems*, 2021.

Gabriel P. Andrade, Rafael M. Frongillo, and Georgios Piliouras. Learning in matrix games can be arbitrarily complex. In *Conference on Learning Theory (COLT)*, 2021.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Neural Information Processing Systems*, 2022.

Eduard Gorbunov, Adrien B. Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *Neural Information Processing Systems*, 2022.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 2015.