

# Perspectives on learning in games

## Tutorial – Part I

Gabriele Farina

MIT

✉ [gfarina@mit.edu](mailto:gfarina@mit.edu)

Athens • 3 July 2024



## Learning in games

*Constructive* answer to the following natural question:

“Can a player that repeatedly plays a game follow rules to refine their strategy after each match, so as to guarantee *mastering* the game in the long run?”

Today, learning-based techniques are typically the fastest way to compute high-quality solutions for large strategic interactions



## Local ↔ global

There exist deep connections between learning and equilibrium. Remarkable:

**Learning:** *dynamic* and *local* (per-player) concept



**Equilibrium:** *static* and *global* (all players) concept

*Global equilibrium emerges from local, decentralized dynamics*

# Goals of this tutorial

---

- Pull together several recent results under a unified point of view that is approachable for newcomers
- Touch on several directions: algorithms, domains, connections between notions, recent trends such as last-iterate convergence and optimistic dynamics
  - I hope this will be useful also for non-newcomers
- Ultimately, some choices had to be made: for example, we will only focus on *discrete* dynamics
- **Part I** focuses mostly on nonsequential, matrix games (normal-form games)
- **Part II** will look at more complicated settings: combinatorial domains and imperfect-information sequential (extensive-form) games

# **Regret and hindsight rationality**

# What does it mean to learn in games?

---

- Philosophical question
- A powerful definition for what “learning in games” means is through the concept of **hindsight rationality**









## Hindsight rationality

*If every single time the player played a certain strategy  $x$ , it would have been strictly better to play a different strategy  $x'$  instead, can we really say that the player has “learnt” how to play...?*

# Formalizing hindsight rationality

---

- To fix ideas, let's look at the case of normal-form games (all the ideas transfer to more general settings)
- Normal-form games are games like rock-paper-scissors:
  - ▶ players choose their strategies simultaneously
  - ▶ only one action is selected and the game ends
  - ▶ given a choice of actions  $(a_i)_{i=1,\dots,n}$  for each player  $i$ , the payoff for each player  $j$  is given by  $u_j(a_1, \dots, a_n)$

			
	0	-1	+1
	+1	0	-1
	-1	+1	0



## Formalizing hindsight rationality

---

- At time  $t$ , every player  $i$  plays according to some strategy  $x_i^{(t)} \in \mathcal{X}_i$ . For rock-paper-scissors, the set of strategies would be the set of all **probability distributions** over the three actions (👊, 🤲, ✂️).

# Formalizing hindsight rationality

---

- At time  $t$ , every player  $i$  plays according to some strategy  $x_i^{(t)} \in \mathcal{X}_i$ . For rock-paper-scissors, the set of strategies would be the set of all **probability distributions** over the three actions (👊, 🤲, ✂️).
- Given the strategies played by the other players, player  $i$ 's expected utility is a **linear function** of their own strategy:  $u_i^{(t)}(x^{(t)}) = \langle u^{(t)}, x^{(t)} \rangle$ .

# Formalizing hindsight rationality

---

- At time  $t$ , every player  $i$  plays according to some strategy  $x_i^{(t)} \in \mathcal{X}_i$ . For rock-paper-scissors, the set of strategies would be the set of all **probability distributions** over the three actions (👊, 🤚, ✌️).
- Given the strategies played by the other players, player  $i$ 's expected utility is a **linear function** of their own strategy:  $u_i^{(t)}(x^{(t)}) = \langle u^{(t)}, x^{(t)} \rangle$ .
- **Idea of hindsight rationality:** Player  $i$  “learnt” to play the game when looking back at the history of play, they cannot think of any transformation  $\varphi : \mathcal{X}_i \rightarrow \mathcal{X}_i$  of their strategies that when applied at the whole history of play would have given strictly better utility

# Input/output model

---

- We assume that each player observes their utility gradient  $u^{(t)}$  as feedback (full information feedback)
  - This can be relaxed to only observing the actual value of the strategy (bandit feedback)
  - Algorithms for bandit feedback typically reduce to the full information case internally by supplying an estimator
- We also assume it's totally fine to output distributions over actions (points in the simplex) instead of specific actions
  - One can always sample one and use concentration arguments to cover the latter case
- The more pressing question is: *what transformations  $\varphi$  are worth considering?*

## Hindsight rationality

Let:

- $\mathcal{X}$  be the set of strategies of the player; and
- $\Phi$  be a set of transformations  $\varphi : \mathcal{X} \rightarrow \mathcal{X}$ .

The  $\Phi$ -regret cumulated up to time  $T$  is given by

$$\Phi \text{ Reg}^{(T)} := \max_{\varphi \in \Phi} \left\{ \sum_{t=1}^T \langle \mathbf{u}^{(t)}, \varphi(\mathbf{x}^{(t)}) \rangle - \langle \mathbf{u}^{(t)}, \mathbf{x}^{(t)} \rangle \right\}$$

## Goal: regret “minimization”

Have  $\Phi \text{ Reg}^{(T)}$  grow sublinearly in  $T$ , *no matter what all the other players do*. Even more generally, sublinear growth no matter the sequence of utility gradients  $\mathbf{u}^{(t)}$

**Some notable choices for the set of transformations  $\Phi$**

# External regret

---

**External** regret:  $\Phi =$  all **constant** functions

$$\varphi_{\hat{x}} : \mathbf{x} \mapsto \hat{x} \quad \forall \mathbf{x} \in \mathcal{X}$$

- Easiest notion: ensure we don't wish we could throw away everything we've played thus far and stick to a single strategy all the times
- So important that it is often called just **regret**
- The definition of  $\Phi$ -regret in this case simplifies into simply

$$\text{Reg}^{(T)} := \max_{\hat{x} \in \mathcal{X}} \left\{ \sum_{t=1}^T \langle \mathbf{u}^{(t)}, \hat{x} \rangle - \langle \mathbf{u}^{(t)}, \mathbf{x}^{(t)} \rangle \right\}.$$

# External regret

---

**External** regret:  $\Phi =$  all **constant** functions



**Warning: (external) regret can be negative!**

This is because we are trading hindsight with the restriction of always using the **same** best action



# External regret

---

**External** regret:  $\Phi =$  all **constant** functions

- In NFGs we can guarantee  $O(\log|A| \sqrt{T})$  external regret
- In convex strategy sets  $\mathcal{X} \subseteq \mathbb{R}^d$ , we can generally guarantee  $O(\text{poly}(d) \sqrt{T})$  external regret
- Connected to coarse correlated equilibria in NFGs and EFGs
  - Special case: Nash eq. in 2-player 0-sum

# External regret

---

**External** regret:  $\Phi =$  all **constant** functions

- In NFGs we can guarantee  $O(\log|A| \sqrt{T})$  external regret
- In convex strategy sets  $\mathcal{X} \subseteq \mathbb{R}^d$ , we can generally guarantee  $O(\text{poly}(d) \sqrt{T})$  external regret
- Connected to coarse correlated equilibria in NFGs and EFGs
  - Special case: Nash eq. in 2-player 0-sum

# External regret

---

**External** regret:  $\Phi =$  all **constant** functions

- In NFGs we can guarantee  $O(\log|A| \sqrt{T})$  external regret
- In convex strategy sets  $\mathcal{X} \subseteq \mathbb{R}^d$ , we can generally guarantee  $O(\text{poly}(d) \sqrt{T})$  external regret
- Connected to coarse correlated equilibria in NFGs and EFGs
  - Special case: Nash eq. in 2-player 0-sum

# Linear swap regret

---

**Linear swap** regret:  $\Phi =$  all **linear functions**  $x \mapsto \mathbf{M}x$

- In NFGs we can guarantee  $O(\text{poly}(|A|)\sqrt{T})$  linear swap regret
- In EFGs we can guarantee  $O(\text{poly}(|E|)\sqrt{T})$  linear swap regret, where  $|E|$  is the number of edges in the (possibly imperfect-information) game tree [FP24]
- Connected to correlated equilibria in NFGs, extensive-form correlated equilibria in EFGs

# Linear swap regret

---

**Linear swap** regret:  $\Phi =$  all **linear functions**  $x \mapsto \mathbf{M}x$

- In NFGs we can guarantee  $O(\text{poly}(|A|)\sqrt{T})$  linear swap regret
- In EFGs we can guarantee  $O(\text{poly}(|E|)\sqrt{T})$  linear swap regret, where  $|E|$  is the number of edges in the (possibly imperfect-information) game tree [FP24]
- Connected to correlated equilibria in NFGs, extensive-form correlated equilibria in EFGs

## Linear swap regret

---

**Linear swap** regret:  $\Phi =$  all **linear functions**  $x \mapsto \mathbf{M}x$

- In NFGs we can guarantee  $O(\text{poly}(|A|)\sqrt{T})$  linear swap regret
- In EFGs we can guarantee  $O(\text{poly}(|E|)\sqrt{T})$  linear swap regret, where  $|E|$  is the number of edges in the (possibly imperfect-information) game tree [FP24]
- Connected to correlated equilibria in NFGs, extensive-form correlated equilibria in EFGs

## More specific transformations $\Phi$

---

The connection with correlated equilibria in NFGs is not a coincidence:

👉 On a probability simplex  $\Delta(A)$ , linear swap transformations in particular include **all probability mass transport**  $\varphi_{a \rightarrow b}$  (for  $a, b \in A, a \neq b$ )

$$\varphi_{a \rightarrow b}(\mathbf{x})[s] := \begin{cases} 0 & \text{if } s = a \quad (\text{remove mass from } a\dots) \\ \mathbf{x}[b] + \mathbf{x}[a] & \text{if } s = b \quad (\dots \text{ and give it to } b) \\ \mathbf{x}[s] & \text{otherwise.} \end{cases}$$

*(“...every time I played 🤔 I should have played 🖐️”...)*

This is known as **internal regret**, and leads to correlated equilibria. It is subsumed by linear swap regret.  $O(\text{polylog}(|A|)\sqrt{T})$  internal regret is possible

# Full swap regret

---

**Full swap regret:  $\Phi =$  all functions  $\mathcal{X} \rightarrow \mathcal{X}$**

- Recent development [Dag+24, PR24]
- In NFGs we can guarantee  $O\left(\log \log |A| \frac{T}{\log T}\right)$  full swap regret
- Note the  $T / \log T$ : *barely* sublinear!
- In EFGs, this is not meaningfully improvable [Das+24]



# Full swap regret

---

**Full swap regret:  $\Phi = \text{all functions } \mathcal{X} \rightarrow \mathcal{X}$**

- Recent development [Dag+24, PR24]
- In NFGs we can guarantee  $O\left(\log \log |A| \frac{T}{\log T}\right)$  full swap regret
- Note the  $T / \log T$ : *barely* sublinear!
- In EFGs, this is not meaningfully improvable [Das+24]

# Full swap regret

---

**Full swap regret:  $\Phi =$  all functions  $\mathcal{X} \rightarrow \mathcal{X}$**

- Recent development [Dag+24, PR24]
- In NFGs we can guarantee  $O\left(\log \log |A| \frac{T}{\log T}\right)$  full swap regret
- Note the  $T / \log T$ : *barely* sublinear!
- In EFGs, this is not meaningfully improvable [Das+24]

# Full swap regret

---

**Full swap regret:  $\Phi =$  all functions  $\mathcal{X} \rightarrow \mathcal{X}$**

- Recent development [Dag+24, PR24]
- In NFGs we can guarantee  $O\left(\log \log |A| \frac{T}{\log T}\right)$  full swap regret
- Note the  $T / \log T$ : *barely* sublinear!
- In EFGs, this is not meaningfully improvable [Das+24]

## More specific transformations $\Phi$

---

- An extremely recent paper (< 2 weeks ago), also made the interesting points that in general one might relax the constraint that each  $\varphi \in \Phi$  map to  $\mathcal{X}$  [Dan+24]
- It is sufficient that  $\varphi \in \Phi$  map to a superset of  $\mathcal{X}$ , as long as  $\varphi$  admits a fixed point  $\varphi(\mathbf{x}^*) = \mathbf{x}^* \in \mathcal{X}$
- These are called by the authors **improper transformations**. There are good reasons to be interested in this relaxation, as it allows for more general notions of regret and a rate-preserving connection to Blackwell's approachability

## **(External) Regret minimization**

External regret minimization is the weakest notion of hindsight rationality we have identified. Hence, one might naturally wonder: why is it so important?

# Why is external regret minimization so important?

---

- Already guarantees best responding\* to a static opponent
- Already leads\* to equilibrium: coarse correlated equilibrium in all convex games
  - Includes NFGs and EFGs
- Special case: leads\* to set of Nash equilibria in two-player zero-sum games
  - This approach has led to superhuman performance in real games
- Surprising breakthrough: with caveats, it is possible to reduce more complicated sets  $\Phi$  to external regret

---

\*: Ergodic convergence. We will talk about last-iterate convergence more at the end

## Application I: Learning a best response against a static opponent

---

- Typically, the difficulty with learning is handling the nonstationarity of the environment, and the fact that everyone is learning at the same time.
- When that is *not* the case, and only one player is learning while nobody else changes their strategy, **it is only reasonable to expect that learning algorithms can learn to best respond**
- This is indeed the case: the average strategy played by the learning player is almost surely a best response to the opponent's strategies

## Application II: Nash equilibrium in two-player zero-sum games

---

- From the definition of regret, it is immediate to check that the product of the average strategies played by external-regret-minimizing players in a two-player zero-sum game is an  $\varepsilon$ -Nash equilibrium, where

$$\varepsilon \leq \frac{\text{Reg}_1^{(T)} + \text{Reg}_2^{(T)}}{T}$$



## Application III: Coarse correlated equilibria

---

- With little extra effort, it can be shown that in *any* general-sum multiplayer game the **average product** distribution of play is an  $\varepsilon$ -coarse correlated equilibrium, where

$$\varepsilon \leq \frac{\max_i \text{Reg}_i^{(T)}}{T}$$

- Note the change from sum to max compared to the two-player zero-sum case

**From external regret to more complicated  $\Phi$  regrets**

## Reducing $\Phi$ -regret to external regret: Gordon et al.'s construction

---

Gordon, Greenwald, and Marks [GGM08] showed that if the following ingredients can be constructed:

- an efficient no-**external**-regret algorithm for  $\Phi$
- an efficient algorithm to compute a fixed point  $x = \varphi(x) \in \mathcal{X}$  of any  $\varphi \in \Phi$

then one can construct an efficient no- $\Phi$ -regret algorithm for  $\mathcal{X}$

## Application: Blum-Mansour's algorithm

---

- As an example of an application of Gordon et al.'s construction (somehow not documented in the literature?), we can recover directly Blum and Mansour's [BM07] algorithm for  $\Phi$ -regret for probability simplexes in the case of

$$\Phi = \text{all } |A| \times |A| \text{ column-stochastic matrices}$$

This is known as **swap regret** and is related to correlated equilibria

# No-external-regret algorithms for simplexes

How can we construct a no-external-regret algorithm for the probability simplex over a finite set of actions?

... We will see algorithms for more complicated sets next time

## Two approaches

---

At a high level, there are two main classes of approaches, both of which are very reasonable:

1. “**Regret tracking style**”: prioritize actions based on the **empirical regret cumulated** so far by them
2. “**Descent style**”: at every iteration **move a little bit in the direction** pointed by the gradient feedback

👉 *Note*: some algorithms, including the important *multiplicative weights update (MWU)*, can be reframed as being part of either category

**“Regret tracking style” algorithms**

## “Regret tracking style” algorithms

---

Let’s start from the **regret tracking style** approaches. At every iteration we keep track of the *empirical regret* cumulated by each of the actions, that is,

$$\mathbf{r}^{(t)}[a] := \sum_{\tau=1}^t \left( \mathbf{u}^{(\tau)}[a] - \langle \mathbf{u}^{(\tau)}, \mathbf{x}^{(\tau)} \rangle \right) \quad \forall a \in A.$$

Conceptually, we want to play more often the actions for which we have highest regret (remember: the regret is the *regret for **not** playing*).



## A first attempt that doesn't work: follow-the-leader

---

- Natural idea: **play the action with the highest cumulated regret**
  - After all, this is the action we wish the most we had played in the past...
  - This algorithm is called *follow-the-leader*
- This idea **does not work**: it jumps around too much
- Consider the sequence of utilities

$$\mathbf{u}^{(1)} = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix}, \quad \mathbf{u}^{(2)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{u}^{(3)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{u}^{(4)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{u}^{(5)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \dots$$

- At every time  $t \geq 2$  we pick the action with value 0
- In hindsight we could have done  $\frac{T}{2}$  by always randomizing the choice, leading to a linear regret

## A first attempt that doesn't work: follow-the-leader

---

It is clear that we need to *soften* the choice of action, so that we don't jump as much. Possible ideas:

## A first attempt that doesn't work: follow-the-leader

---

It is clear that we need to *soften* the choice of action, so that we don't jump as much. Possible ideas:

- We can replace picking the action with the highest regret with picking actions **proportionally** to their regret → [regret matching](#)

## A first attempt that doesn't work: follow-the-leader

---

It is clear that we need to *soften* the choice of action, so that we don't jump as much. Possible ideas:

- We can replace picking the action with the highest regret with picking actions **proportionally** to their regret → **regret matching**
- We can use the softmax function for some finite temperature → **multiplicative weights update**

## A first attempt that doesn't work: follow-the-leader

---

It is clear that we need to *soften* the choice of action, so that we don't jump as much. Possible ideas:

- We can replace picking the action with the highest regret with picking actions **proportionally** to their regret → regret matching
- We can use the softmax function for some finite temperature → multiplicative weights update
- We can in general *regularize* the choice → follow-the-regularized-leader

All of these ideas work.

## Regret matching

---

We cannot pick the action with the highest regret, but can we pick an action **proportionally** to their regret?

## Regret matching

---

We cannot pick the action with the highest regret, but can we pick an action **proportionally** to their regret?

... almost. What about actions with negative regret?

## Regret matching

---

We cannot pick the action with the highest regret, but can we pick an action **proportionally** to their regret?

... almost. What about actions with negative regret? *We ignore those!*



# Regret matching

---

We cannot pick the action with the highest regret, but can we pick an action **proportionally** to their regret?

... almost. What about actions with negative regret?



## Regret matching idea

Pick an action **proportionally** to the “ReLU” of their empirical regret

$$\mathbf{x}^{(t)} \propto [\mathbf{r}^{(t)}]^+.$$

(If the right-hand-side is zero, pick uniformly at random)

# Regret matching analysis

---

- Main idea:  $\mathbf{r}^{(t+1)} - \mathbf{r}^{(t)} \perp \mathbf{x}^{(t)}$  (this is always true, not just for regret matching)
- In regret matching,  $\mathbf{x}^{(t)} \propto [\mathbf{r}^{(t)}]^+$  so in particular

$$\mathbf{r}^{(t+1)} - \mathbf{r}^{(t)} \perp [\mathbf{r}^{(t)}]^+$$

(this is trivially true even in the edge case in which  $\mathbf{r}^{(t)} = \mathbf{0}$ )

- Now use the inequality

$$\|[\mathbf{a} + \mathbf{b}]^+\|_2^2 \leq \|[\mathbf{a}]^+ + \mathbf{b}\|_2^2$$

applied to  $\mathbf{a} = \mathbf{r}^{(t)}$  and  $\mathbf{b} = \mathbf{r}^{(t+1)} - \mathbf{r}^{(t)}$

# Regret matching analysis

---

- We immediately obtain

$$\left\| [\mathbf{r}^{(t+1)}]^+ \right\|_2^2 \leq \left\| [\mathbf{r}^{(t)}]^+ + (\mathbf{r}^{(t+1)} - \mathbf{r}^{(t)}) \right\|_2^2 = \left\| [\mathbf{r}^{(t)}]^+ \right\|_2^2 + \left\| \mathbf{r}^{(t+1)} - \mathbf{r}^{(t)} \right\|_2^2$$

- Hence by induction we have

$$\left\| [\mathbf{r}^{(T)}]^+ \right\|_2^2 \leq \sum_{t=1}^T \left\| \mathbf{u}^{(t)} - \langle \mathbf{u}^{(t)}, \mathbf{x}^{(t)} \rangle \mathbf{1} \right\|_2^2$$

- Taking square roots, this shows  $\text{Reg}^{(T)} \leq \sqrt{|A|} T$  assuming the utilities are all in  $[-1, 1]$

# Multiplicative Weights Update

---

## MWU

Another approach is to smoothen out the hard argmax we were operating in follow-the-leader, and replace it with a **softmax**:

$$\mathbf{x}^{(t+1)} \propto \exp(\eta \mathbf{r}^{(t)})$$

where  $\eta > 0$  is the learning rate, and the exponentiation is component-wise

This algorithm is called Multiplicative Weights Update (MWU)

# Follow-the-regularized-leader

---

- All the algorithms we have seen so far are directly or indirectly instances of a meta-algorithm called **follow-the-regularized-leader (FTRL)**
  - Not obvious for regret matching [\[FKS21\]](#)

## FTRL

Let  $\psi : \mathcal{X} \rightarrow \mathbb{R}$  be 1-strongly convex wrt some norm  $\|\cdot\|$ . At every time  $t$ , FTRL produces the strategy

$$\mathbf{x}^{(t+1)} := \arg \max_{\mathbf{x} \in \Delta(A)} \left\{ \langle \mathbf{r}^{(t)}, \mathbf{x} \rangle - \frac{1}{\eta} \psi(\mathbf{x}) \right\}$$

## Aside: Follow-the-regularized-leader on convex & closed sets

---

👉 *Aside:* FTRL is far more general than probability simplexes  $\Delta(A)$ , though in general the vector of regrets must be replaced with the **cumulative utility**

### FTRL

Let  $\mathcal{X}$  be a convex and closed set and  $\psi : \mathcal{X} \rightarrow \mathbb{R}$  be 1-strongly convex wrt some norm  $\|\cdot\|$ . At every time  $t$ , the general form of FTRL produces the point

$$\mathbf{x}^{(t+1)} := \arg \max_{\mathbf{x} \in \mathcal{X}} \left\{ \left\langle \sum_{\tau=1}^t \mathbf{u}^{(\tau)}, \mathbf{x} \right\rangle - \frac{1}{\eta} \psi(\mathbf{x}) \right\}$$

# Follow-the-regularized-leader

---

## Theorem

No matter the sequence of utilities  $\mathbf{u}^{(t)}$  received by FTRL, the strategies  $\mathbf{x}^{(t)}$  produced by FTRL satisfy the regret bound

$$\text{Reg}^{(T)}(\mathbf{x}^*) \leq \frac{D_\psi(\mathbf{x}^* \parallel \mathbf{x}^{(1)})}{\eta} + \eta \sum_{t=1}^T \|\mathbf{u}^{(t)}\|_*^2 - \frac{1}{8\eta} \sum_{t=2}^T \|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|^2$$

where  $\|\cdot\|_*$  is the norm dual to  $\|\cdot\|$  and  $D_\psi$  denotes the Bregman divergence induced by  $\psi$ .

Proof: special case of a more general result for **predictive** FTRL we will see later.

## MWU as FTRL

---

- Multiplicative Weights Update is simply FTRL where  $\psi$  is chosen to be the **negative entropy** function

$$\psi(\mathbf{x}) = - \sum_{a \in A} \mathbf{x}[a] \log \mathbf{x}[a]$$

which is 1-strongly convex wrt the  $l_1$  norm

- The Bregman divergence induced by  $\psi$  is the KL divergence. Note that  $\mathbf{x}^{(1)}$  is uniform and therefore

$$D_{\psi}(\mathbf{x}^* \parallel \mathbf{x}^{(1)}) \leq \log|A|$$

for all  $\mathbf{x}^* \in \Delta(A)$ .



# MWU as FTRL

---

- Hence,

$$\begin{aligned}\text{Reg}^{(T)}(\mathbf{x}^*) &\leq \frac{D_\psi(\mathbf{x}^* \parallel \mathbf{x}^{(1)})}{\eta} + \eta \sum_{t=1}^T \|\mathbf{u}^{(t)}\|_*^2 - \frac{1}{\eta} \sum_{t=2}^T \|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|^2 \\ &\leq \frac{\log|A|}{\eta} + \eta \sum_{t=1}^T \|\mathbf{u}^{(t)}\|_\infty^2.\end{aligned}$$

- Assuming that all utilities of the game are in  $[-1, 1]$  and picking  $\eta = \frac{\sqrt{\log|A|}}{\sqrt{T}}$ , we find that

$$\text{Reg}^{(T)} \leq \sqrt{\log|A|} T,$$

## MWU vs Regret Matching

---

- In **theory**, MWU has a better regret bound than regret matching
  - $O(\sqrt{\log|A| T})$  vs  $O(\sqrt{|A| T})$
- In **practice**, regret matching is preferred because of its lack of hyperparameters (learning rate) to tune

**“Descent style” algorithms**

## “Descent style” algorithms

---

- Another class of algorithms is based on the idea of online gradient descent: at every iteration **nudge the strategy in the direction of the given utility gradient**

### Online projected gradient descent (OPGD)

Let  $\mathcal{X}$  be closed and convex. At every time  $t$ , OPGD picks the strategy

$$\mathbf{x}^{(t+1)} := \text{Proj}_{\mathcal{X}}(\mathbf{x}^{(t)} + \eta \mathbf{u}^{(t)}) = \arg \min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - (\mathbf{x}^{(t)} + \eta \mathbf{u}^{(t)})\|_2^2.$$

# Online mirror descent

---

- Akin to what happens in offline optimization, we can generalize (online) projected gradient descent into (online) mirror descent.

## Online mirror descent (OMD)

Let  $\mathcal{X}$  be closed and convex and  $\psi$  be 1-strongly convex with respect to some norm  $\|\cdot\|$ . At every time  $t$ , OMD picks the strategy

$$\mathbf{x}^{(t+1)} := \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{u}^{(t)}, \mathbf{x} \rangle - \frac{1}{\eta} D_{\psi}(\mathbf{x} \parallel \mathbf{x}^{(t)}) \right\}.$$

# Online mirror descent

---

## Theorem

No matter the sequence of utilities  $\mathbf{u}^{(t)}$  received by OMD, the strategies  $\mathbf{x}^{(t)}$  produced by OMD satisfy the regret bound

$$\text{Reg}^{(T)}(\mathbf{x}^*) \leq \frac{D_\psi(\mathbf{x}^* \parallel \mathbf{x}^{(1)})}{\eta} + \eta \sum_{t=1}^T \|\mathbf{u}^{(t)}\|_*^2 - \frac{1}{8\eta} \sum_{t=2}^T \|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|^2$$

where  $\|\cdot\|_*$  is the norm dual to  $\|\cdot\|$  and  $D_\psi$  denotes the Bregman divergence induced by  $\psi$ .

Proof: special case of a more general result for **predictive** OMD we will see later.

## MWU as OMD

---

- It turns out that MWU is also an instance of OMD, not just FTRL!
  - For the same  $\psi$  set as **negative entropy** we used before
- In fact, OMD = FTRL whenever  $\psi$  is Legendre, which means that the gradients of  $\psi$  explode at the boundary of the strategy set  $\mathcal{X}$

# **Optimism and predictivity**



# What is optimism

---

In recent years, there has been a lot of interest in the idea of **optimism** in learning algorithms



**Nonstationary  $\neq$  adversarial**

When all players learn at the same time, the environment is stationary but not necessarily adversarial

Can take advantage of this to design learning algorithms with better regret guarantees and convergence properties?

# Predictivity and Optimism

---



## Predictivity

The idea of optimism is to **anticipate** the next utility gradient  $\mathbf{u}^{(t+1)}$  by having a **prediction**  $\mathbf{m}^{(t+1)}$



## Optimism

The idea of optimism is to use predictivity with the specific guess  $\mathbf{m}^{(t+1)} = \mathbf{u}^{(t)}$  at all times  $t$ . This corresponds to predicting that the feedback is slow-changing

# What is optimism

---

- Standard FTRL:

$$\mathbf{x}^{(t+1)} := \arg \max_{\mathbf{x} \in \mathcal{X}} \left\{ \left\langle \sum_{\tau=1}^t \mathbf{u}^{(\tau)}, \mathbf{x} \right\rangle - \frac{1}{\eta} \psi(\mathbf{x}) \right\}$$

- **Predictive FTRL:**

$$\mathbf{x}^{(t+1)} := \arg \max_{\mathbf{x} \in \mathcal{X}} \left\{ \left\langle \mathbf{m}^{(t+1)} + \sum_{\tau=1}^t \mathbf{u}^{(\tau)}, \mathbf{x} \right\rangle - \frac{1}{\eta} \psi(\mathbf{x}) \right\}$$

- **Optimistic FTRL:**

$$\mathbf{x}^{(t+1)} := \arg \max_{\mathbf{x} \in \mathcal{X}} \left\{ \left\langle \mathbf{u}^{(t)} + \sum_{\tau=1}^t \mathbf{u}^{(\tau)}, \mathbf{x} \right\rangle - \frac{1}{\eta} \psi(\mathbf{x}) \right\}$$

# What is optimism

---

- Standard OMD:

$$\mathbf{x}^{(t+1)} := \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{u}^{(t)}, \mathbf{x} \rangle - \frac{1}{\eta} D_{\psi}(\mathbf{x} \parallel \mathbf{x}^{(t)}) \right\}$$

- **Predictive OMD:**

$$\mathbf{x}^{(t+1)} := \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{u}^{(t)} + (\mathbf{m}^{(t+1)} - \mathbf{m}^{(t)}), \mathbf{x} \rangle - \frac{1}{\eta} D_{\psi}(\mathbf{x} \parallel \mathbf{x}^{(t)}) \right\}$$

- **Optimistic OMD:**

$$\mathbf{x}^{(t+1)} := \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{u}^{(t)} + (\mathbf{u}^{(t)} - \mathbf{u}^{(t-1)}), \mathbf{x} \rangle - \frac{1}{\eta} D_{\psi}(\mathbf{x} \parallel \mathbf{x}^{(t)}) \right\}$$

“Vanilla” FTRL and OMD correspond to the case where the prediction is always set to zero:  $\mathbf{m}^{(t)} = \mathbf{0}$  for all  $t$

# RVU bounds

---

## Theorem

Predictive FTRL and predictive OMD satisfy the regret bound

$$\text{Reg}^{(T)}(\mathbf{x}^*) \leq \frac{D_\psi(\mathbf{x}^* \parallel \mathbf{x}^{(1)})}{\eta} + \eta \sum_{t=1}^T \|\mathbf{u}^{(t)} - \mathbf{m}^{(t)}\|_*^2 - \frac{1}{8\eta} \sum_{t=2}^T \|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|^2$$

- The above form of the regret bound is called an **RVU bound** (regret bounded by variation in utilities) [Syr+15]
- RVU bounds come in useful for a variety of purposes

## Accelerated convergence to two-player zero-sum Nash

---

- When all players use optimistic FTRL or optimistic OMD in a two-player zero-sum game, the average strategies converge to the set of Nash equilibria at a rate of  $O(\frac{1}{T})$  [Syr+15]
- We know from before that the sum of the regrets of the players bounds the Nash equilibrium approximation

## Accelerated convergence to two-player zero-sum Nash

---

- In a two-player zero-sum game, the utility of player 1 is given by  $u_1^{(t)} = \mathbf{A}\mathbf{y}^{(t)}$  and that of player 2 by  $u_2^{(t)} = \mathbf{A}^\top \mathbf{x}^{(t)}$ , leading to

$$\text{Reg}_1^{(T)} \leq \frac{O_T(1)}{\eta} + \eta \sum_{t=1}^T \|\mathbf{A}(\mathbf{y}^{(t)} - \mathbf{y}^{(t-1)})\|_*^2 - \frac{1}{8\eta} \sum_{t=2}^T \|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|^2$$

$$\text{Reg}_2^{(T)} \leq \frac{O_T(1)}{\eta} + \eta \sum_{t=1}^T \|\mathbf{A}^\top (\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)})\|_*^2 - \frac{1}{8\eta} \sum_{t=2}^T \|\mathbf{y}^{(t)} - \mathbf{y}^{(t-1)}\|^2$$

- Summing and picking  $\eta$  constant small enough (which depends on the operator norm of  $\mathbf{A}$  and the norm  $\|\cdot\|$ ), we obtain

$$\text{Reg}_1^{(T)} + \text{Reg}_2^{(T)} = O_T(1)$$



## Accelerated convergence to equilibrium: the general case

---

- Rates of  $\tilde{O}\left(\frac{1}{T}\right)$  for coarse correlated and correlated equilibria (CCE) in normal-form games (and beyond) are also known for the multiplayer case, but they are much harder to prove
- One of the main obstacles: convergence to CCE is driven by the **max** of the regrets of the players, not the sum
- [Syr+15] showed  $O\left(n \log|A| T^{-\frac{3}{4}}\right)$  for OMWU using RVU bounds
  - Improved by [CP20] to  $O\left(n \log^{\frac{5}{6}}|A| T^{-\frac{5}{6}}\right)$  for two-player general-sum games only

## Accelerated convergence to equilibrium: the general case

---

- [DFG21] showed  $O\left(n \log|A| \frac{\log^4 T}{T}\right)$  convergence for OMWU using a very complicated analysis based on the idea of high-order stability
- [Far+22] showed  $O\left(n |A| \frac{\log T}{T}\right)$  using RVU bounds paired with a special regularizer
- [FPS24] Under submission:  $O\left(n \log^2 |A| \frac{\log T}{T}\right)$  using RVU bounds + nonmonotonic learning rate control
- Can  $O\left(n \log|A| \frac{1}{T}\right)$  be achieved in general games?

**The question of convergence in iterates**

Most guarantees of learning in games are possessed by the *average* distribution of play



### Iterate convergence

This is a bit unsatisfactory: can we guarantee that the *last* iterate or *best* iterate of the dynamics converges to a good solution?

## Iterate convergence

---

- Complexity theoretic considerations imply that iterate convergence is not possible beyond two-player zero-sum games
- In two-player zero-sum games, what is known?
- Most strong results (that is, with good rates) are known only for *optimistic* online projected gradient descent
- Asymptotic results (*i.e.*, without rates) are also known for *optimistic* multiplicative weights update

# Optimistic gradient descent-ascent

---

- There is a short and sweet proof that optimistic OPGD has  $O\left(\frac{1}{\sqrt{T}}\right)$  **best-iterate** convergence to the set of Nash equilibria in two-player zero-sum games
  - Proof from [\[Ana+22\]](#)
- The idea is to use the fact that the sum of the regrets of the two players must be nonnegative together with the RVU bounds of the players

## Optimistic gradient descent-ascent

---

- What about **last**-iterate convergence?
- The result holds, but it is significantly more complicated to prove [COZ22]
- Proof uses a Lyapunov function that was constructed using SOS programming
- Confirms a bound of  $O_T\left(\frac{1}{\sqrt{T}}\right)$  convergence, hiding only polynomial constants in the number of actions for any payoff matrix with entries in  $[-1, 1]$

# What about optimistic multiplicative weights update?

---

- As we have seen so far, optimistic MWU has some of the properties known:
  - $\log|A|$  dependence on the number of actions  $A$
  - $\log^4 T$  dependence on the number of iterations  $T$
- It is known that in two-player zero-sum games, optimistic MWU has **asymptotic** convergence to the set of Nash equilibria [DP19, HAM21]
- If the Nash is unique, then the convergence is linear, but tainted by possibly exponentially large problem-dependent constants [Wei+20] (similar to Tseng-type analysis of extragradient [Tse95])
- **What about concrete rates?**



I believe it's fair to say that many people in the field believed that good last-iterate convergence of OMWU was just one good trick away

After all, OMWU has always spoiled us with its good properties

## FTRL struggles with last-iterate convergence

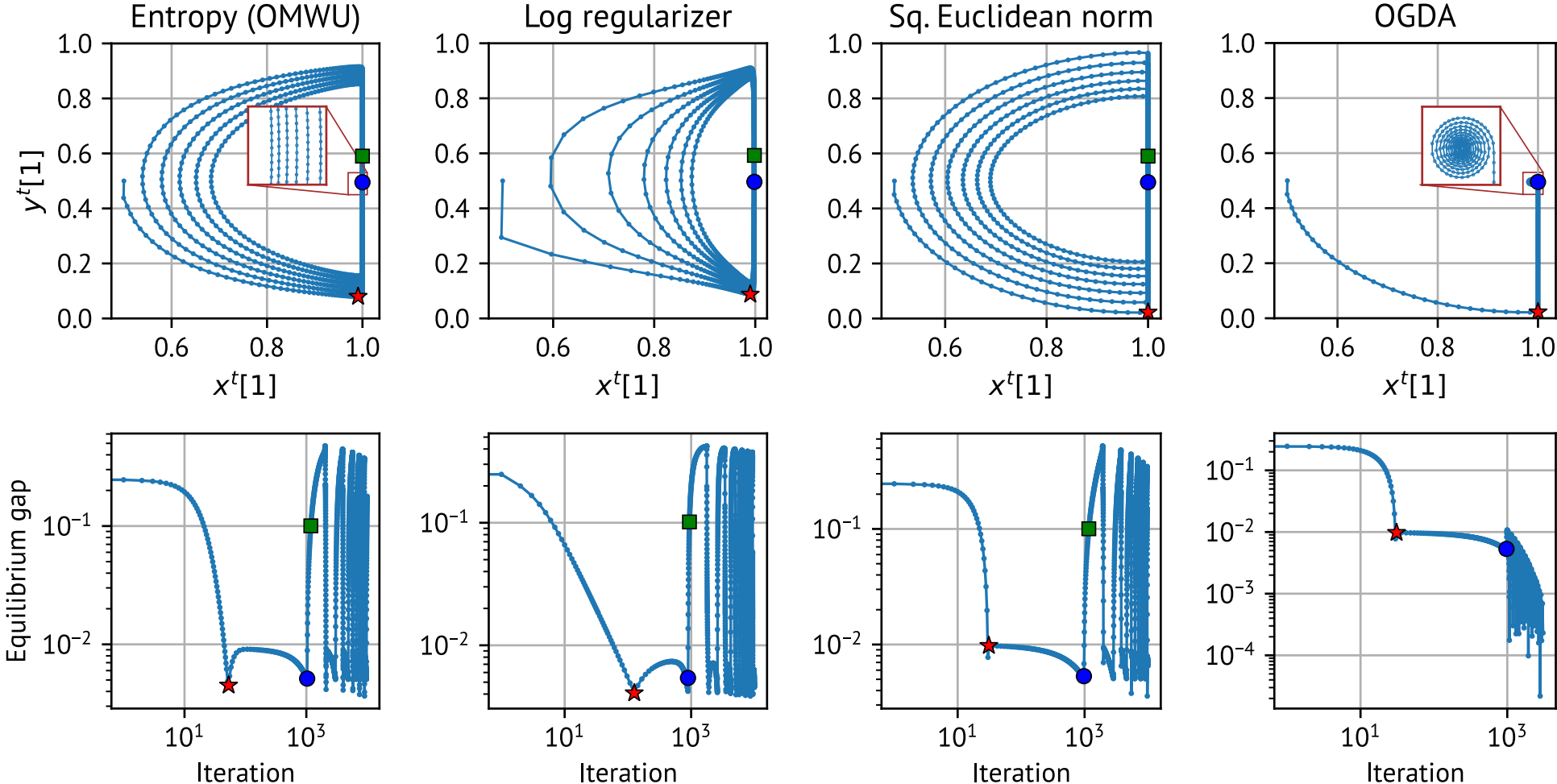
---

- It turns out things are not like that: in a very recent development, it was shown that OMWU cannot possibly converge in iterates (neither best nor last) [Cai+24]
- In fact, the lack of last-iterate convergence applies to any known instance of FTRL
- Constructive proof by analyzing the dynamics in the same  $2 \times 2$  game

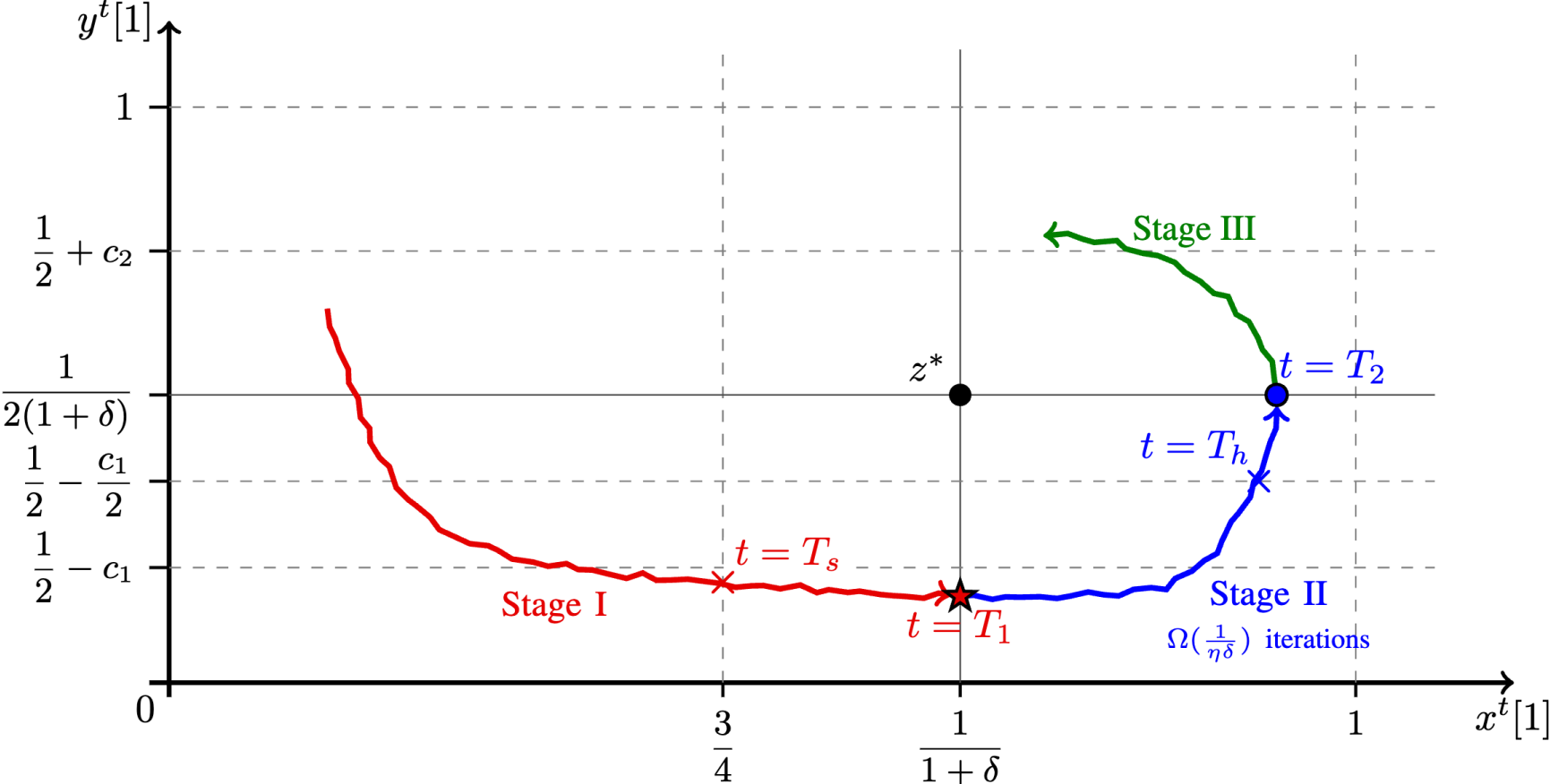
$$A_\delta := \begin{pmatrix} \frac{1}{2} + \delta & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$$

whose unique Nash equilibrium is at  $x^* := \left( \frac{1}{1+\delta}, \frac{\delta}{1+\delta} \right), y^* := \left( \frac{1}{2(1+\delta)}, \frac{1+2\delta}{2(1+\delta)} \right)$ .

# FTRL struggles with last-iterate convergence



# FTRL struggles with last-iterate convergence



# FTRL struggles with last-iterate convergence

---

## Theorem

Under standard assumptions about the regularizer, there is no function  $f$  such that optimistic FTRL produces a last-iterate convergence rate of  $f(d_1, d_2, T) \rightarrow 0$  where the entries of the loss matrix are in  $[0, 1]$ , and  $d_1$  and  $d_2$  are the number of actions of the players

# Bibliography

- [FP24] G. Farina and C. Pipis, “Polynomial-time linear-swap regret minimization in imperfect-information sequential games,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [Dag+24] Y. Dagan, C. Daskalakis, M. Fishelson, and N. Golowich, “From External to Swap Regret 2.0: An Efficient Reduction for Large Action Spaces,” in *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, 2024, pp. 1216–1222.
- [PR24] B. Peng and A. Rubinstein, “Fast swap regret minimization and applications to approximate correlated equilibria,” in *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, 2024, pp. 1223–1234.

- [Das+24] C. Daskalakis, G. Farina, N. Golowich, T. Sandholm, and B. H. Zhang, “A Lower Bound on Swap Regret in Extensive-Form Games,” *arXiv preprint arXiv:2406.13116*, 2024.
- [Dan+24] C. Dann, Y. Mansour, M. Mohri, J. Schneider, and B. Sivan, “Rate-Preserving Reductions for Blackwell Approachability,” *arXiv preprint arXiv:2406.07585*, 2024.
- [GGM08] G. J. Gordon, A. Greenwald, and C. Marks, “No-regret learning in convex games,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 360–367.
- [BM07] A. Blum and Y. Mansour, “From external to internal regret.,” *Journal of Machine Learning Research*, vol. 8, no. 6, 2007.

- [FKS21] G. Farina, C. Kroer, and T. Sandholm, “Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 5363–5371.
- [Syr+15] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire, “Fast convergence of regularized learning in games,” *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [CP20] X. Chen and B. Peng, “Hedging in games: Faster convergence of external and swap regrets,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 18990–18999, 2020.
- [DFG21] C. Daskalakis, M. Fishelson, and N. Golowich, “Near-optimal no-regret learning in general games,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27604–27616, 2021.



- [Far+22] G. Farina, I. Anagnostides, H. Luo, C.-W. Lee, C. Kroer, and T. Sandholm, “Near-optimal no-regret learning dynamics for general convex games,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 39076–39089, 2022.
- [Ana+22] I. Anagnostides, I. Panageas, G. Farina, and T. Sandholm, “On last-iterate convergence beyond zero-sum games,” in *International Conference on Machine Learning*, 2022, pp. 536–581.
- [COZ22] Y. Cai, A. Oikonomou, and W. Zheng, “Tight Last-Iterate Convergence of the Extragradient and the Optimistic Gradient Descent-Ascent Algorithm for Constrained Monotone Variational Inequalities,” in *NeurIPS*, 2022.
- [DP19] C. Daskalakis and I. Panageas, “Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization,” in *Innovations in Theoretical Computer Science*, 2019.

- [HAM21] Y.-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos, “Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium,” in *Conference on Learning Theory*, 2021, pp. 2388–2422.
- [Wei+20] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo, “Linear Last-iterate Convergence in Constrained Saddle-point Optimization,” in *International Conference on Learning Representations*, 2020.
- [Tse95] P. Tseng, “On linear convergence of iterative methods for the variational inequality problem,” *Journal of Computational and Applied Mathematics*, vol. 60, no. 1–2, pp. 237–252, 1995.
- [Cai+24] Y. Cai *et al.*, “Fast Last-Iterate Convergence of Learning in Games Requires Forgetful Algorithms,” *arXiv preprint arXiv:2406.10631*, 2024.