

Faster No-Regret Learning Dynamics and Last-Iterate Convergence



15 888 **Computational Game Solving** (Fall 2026)
Ioannis Anagnostides

Today's lecture

- Near-optimal regret in self-play
 - Optimistic regret minimization
 - The RVU property
 - Zero-sum and general-sum games
- Last-iterate convergence
 - Connections to optimistic regret minimization
 - Connections to price of anarchy and smooth games

Regret minimization against an adversary vs. self-play

- We have seen that players with no-regret converge to equilibria
- The rate of convergence is driven by their regrets
- What's the best rate we can hope for?
- The adversarial setting is *overly pessimistic*; when learning in games, we have control over the sequence of utilities
- Can we improve our analysis?



Barriers with traditional learning algorithms

Theorem. For any learning rate, when both players in a two-player game employ MWU, at least one of the players will have $\sqrt{T}$ regret



We need new algorithmic ideas!

- Similar lower bounds are known for other common algorithms, such as RM and RM+

Optimistic regret minimization

Optimistic FTRL

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

Optimistic mirror descent

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

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



The only difference is that we have a **prediction** vector, typically set as the *previously observed utility*

Regret bounded by variation in utilities

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

Two key differences with the usual regret bound:

- The regret is bounded by the *misprediction error*
- There is a negative term that *decreases* the regret when the player is changing its strategies rapidly (!)

Theorem. Both OFTRL and OMD satisfy the RVU bound

Predictive regret matching

Algorithm 1: Predictive RM (PRM)

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

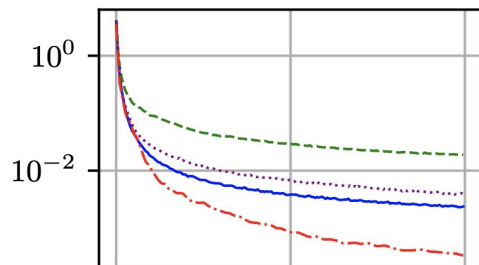
Algorithm 2: Predictive RM⁺ (RM⁺)

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

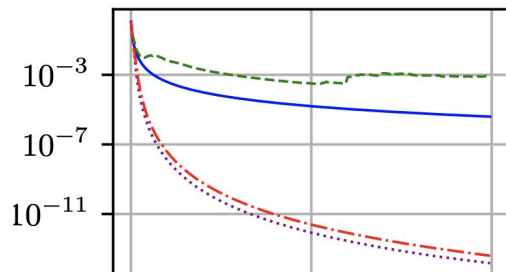
PRM in action

From Farina et al.

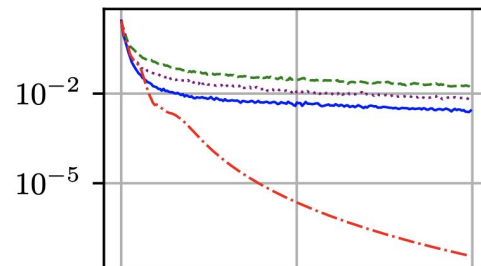
[A] Goofspiel



[B] Liar's dice



[C] Battleship



--- PCFR⁺ — CFR⁺
--- LCFR DCFR

Faster rates using stability

Regularized algorithms, such as (O)FTRL and (O)MD guarantee

$$\|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\| \leq O(\eta).$$

- Two consecutive strategies do not change by a lot
- Does *not* hold for regret matching

$$\|\mathbf{u}_i^{(t)} - \mathbf{u}_i^{(t-1)}\|_\infty \leq \sum_{i' \neq i} \|\mathbf{x}_{i'}^{(t)} - \mathbf{x}_{i'}^{(t-1)}\|_1, \text{ where } \mathbf{u}_i^{(t)} = \mathbf{u}_i(\mathbf{x}_{-i}^{(t)}).$$

- If all players employ regularized algorithms, the *utilities are changing slowly*
- The utility is a polynomial (by expanding the expectation), so it's also Lipschitz continuous in the strategies

Faster rates using stability

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

- Optimizing the learning rate, the regret is bounded by $T^{1/4}$
- This is still far from the lower bound
- Can we do better?

The sum of the regrets is bounded

$$\begin{aligned}\sum_{i=1}^n \text{Reg}_i^{(T)} &\leq \alpha n + (n-1)\beta \sum_{i=1}^n \sum_{i' \neq i} \sum_{t=1}^T \|\mathbf{x}_{i'}^{(t)} - \mathbf{x}_{i'}^{(t)}\|_1^2 - \gamma \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}\|_1^2 \\ &\leq \alpha n + (n-1)^2\beta \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t)}\|_1^2 - \gamma \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}\|_1^2 \\ &\leq \alpha n + \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}\|_1^2 ((n-1)^2\beta - \gamma) \\ &\leq \alpha n - \frac{\gamma}{2} \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}\|_1^2,\end{aligned}$$



We care about the **maximum** of the regrets

Games with nonnegative sum of regrets

Theorem. For any game with nonnegative sum of regrets,

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

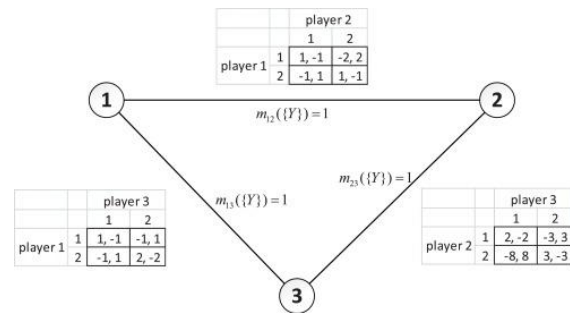
The misprediction error is bounded!

The assumption holds for zero-sum games:

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

Polymatrix zero-sum games

- A generalization of two-player zero-sum games
- There is a graph, and every player is uniquely associated with a node
- Every edge represents a (two-player) zero-sum game between the incident players
- A player gets the sum of the utilities from all the individual games



Taken from Deng et al.

Extending to general-sum games

- The previous argument only applies to games with nonnegative sum of regrets, which is a severe restriction
- How can we extend it to general-sum games?
- What if we consider instead a **nonnegative measure of regret**?



Swap regret is nonnegative!



It suffices to prove an RVU bound for swap regret

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

A reminder of Blum-Mansour

Algorithm 2: Blum-Mansour algorithm for minimizing swap regret

```
1 Input: A regret minimizer  $\mathfrak{R}_a$  for each action  $a \in \mathcal{A}$ 
2 NEXTSTRATEGY():
3   for each action  $a \in \mathcal{A}$  do
4      $\Delta(\mathcal{A}) \ni \mathbf{x}_a^{(t)} := \mathfrak{R}_a.\text{NEXTSTRATEGY}();$ 
5   Set  $\mathbf{M}^{(t)} := [(\mathbf{x}_a^{(t)})_{a \in \mathcal{A}}]$  ;
6   return  $\Delta(\mathcal{A}) \ni \mathbf{x}^{(t)} = \mathbf{M}^{(t)} \mathbf{x}^{(t)};$ 

7 OBSERVEUTILITY( $\mathbf{u}^{(t)} \in \mathbb{R}^{\mathcal{A}}$ ):
8   for each action  $a \in \mathcal{A}$  do
9     Set  $\mathbf{u}_a^{(t)} := \mathbf{x}^{(t)}[a]\mathbf{u}^{(t)}$  ;
10     $\mathfrak{R}_a.\text{OBSERVEUTILITY}(\mathbf{u}_a^{(t)});$ 
```

RVU bound for Blum-Mansour

We can use the RVU bound for each individual local regret minimizer

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

It suffices to prove

$$\|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|_1 \leq C \sum_{a \in \mathcal{A}} \|\mathbf{x}_a^{(t)} - \mathbf{x}_a^{(t-1)}\|$$

Stability of fixed points

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

Stability of fixed points

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

- Those two Markov chains are close to each other in terms of transition probs
- But their stationary distributions are not!

Multiplicative stability

We need a more refined notion of stability—**multiplicative stability**

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

- The ratio of two consecutive coordinates has to be close to 1
- In the previous example, the stochastic matrices are **not** multiplicatively close
- Most algorithms we have seen do *not* guarantee this notion of stability; *but MWU does*

Stability of fixed points

Theorem. If the transition probabilities of two Markov chains are multiplicatively close, their fixed points will also be close.

Proof by Markov chain tree theorem:

$$\mathbf{x}[a] = \frac{\sum_{\mathcal{T} \in \mathbb{T}_a} \prod_{(u,v) \in E(\mathcal{T})} \mathbf{M}[v, u]}{\sum_{a \in \mathcal{A}} \sum_{\mathcal{T} \in \mathbb{T}_a} \prod_{(u,v) \in E(\mathcal{T})} \mathbf{M}[v, u]}.$$

👉 Stability ensures we can get a bound of $T^{1/4}$

Improved regularizer

The second key idea is to use the **logarithmic regularizer**:

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

Adds a $\log T$
dependence

- The range is unbounded, but can be handled by pushing the comparator away from the boundary
- The main benefit is that we get a refined **local norm**, which is dynamically changing over time

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

RVU for swap regret

Using the Markov chain tree theorem, $\|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|_1 \leq C \sum_{a \in \mathcal{A}} \|\mathbf{x}_a^{(t)} - \mathbf{x}_a^{(t-1)}\|_{\mathbf{x}_a^{(t-1)}}.$

Theorem. There is an algorithm that satisfies the RVU bound with respect to swap regret.

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

As a result, we can guarantee that every player in a general-sum game will have **logarithmic regret**, which is near-optimal.

Adversarial robustness

- What if one or more players deviate from the protocol?
- Can we still get the best regret possible when facing an adversary?

Adversarial robustness

- What if one or more players deviate from the protocol?
- Can we still get the best regret possible when facing an adversary?
- It's enough to keep track of the misprediction error

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



If it gets larger than logarithmic, we can switch to an algorithm tuned for the adversarial regime

Near-optimal variant of PRM+

Theorem. PRM and PRM+ can have $\sqrt{T}$ regret even in two-player zero-sum games.

- This is in contrast to OFTRL and OMD
- The fundamental problem with PRM and PRM+ is **instability**
- There's a fix!
- The key idea is to make sure that the norm of the *regret vector is increasing*
 - The bad case arises when the regret vector is close to zero

Increasing regret (IR) PRM+

Algorithm 2: IR-PRM and IR-PRM⁺

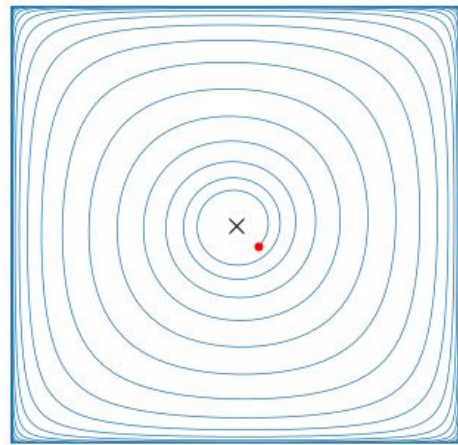
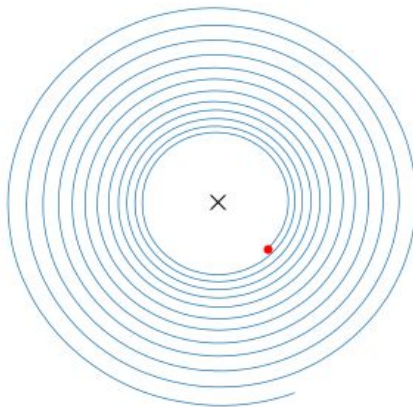
```
1 function INITIALIZE()
2    $\tilde{\mathbf{r}}^{(1)} \leftarrow$  arbitrary vector in  $\mathbb{R}_{\geq 0}^n$ 
3    $\tilde{\mathbf{x}}^{(1)} \leftarrow \tilde{\mathbf{r}}^{(1)} / \|\tilde{\mathbf{r}}^{(1)}\|_1$ 
4 function NEXTSTRATEGY(prediction  $\mathbf{m}^{(t)} \in \mathbb{R}^n$ )
5   if  $[\tilde{\mathbf{r}}^{(t)}]_+ = \mathbf{0}$  then
6      $\mathbf{m}^{(t)} \leftarrow \mathbf{0}$ 
7     return  $\mathbf{x}^{(t)} \leftarrow \tilde{\mathbf{x}}^{(t)}$ 
8   let  $\gamma \in \mathbb{R}$  be s.t.  $\|[\tilde{\mathbf{r}}^{(t)} + \mathbf{m}^{(t)} - \gamma \mathbf{1}]_+\|_2 = \|\tilde{\mathbf{r}}^{(t)}\|_2$ 
9    $\mathbf{r}^{(t)} \leftarrow \tilde{\mathbf{r}}^{(t)} + \mathbf{m}^{(t)} - \gamma \mathbf{1}$ 
10  return  $\mathbf{x}^{(t)} \leftarrow [\mathbf{r}^{(t)}]_+ / \|\mathbf{r}^{(t)}\|_1$ 
11 function OBSERVEUTILITY(utility  $\mathbf{u}^{(t)} \in \mathbb{R}^n$ )
12   $\mathbf{g}^{(t)} \leftarrow \mathbf{u}^{(t)} - \mathbf{m}^{(t)} - \langle \mathbf{u}^{(t)} - \mathbf{m}^{(t)}, \mathbf{x}^{(t)} \rangle$ 
13   $\tilde{\mathbf{r}}^{(t+1)} \leftarrow [\mathbf{r}^{(t)} + \mathbf{g}^{(t)}]_+$ 
14   $\tilde{\mathbf{x}}^{(t+1)} \leftarrow \tilde{\mathbf{r}}^{(t+1)} / \|\tilde{\mathbf{r}}^{(t+1)}\|_1$ 
```

Zhang et al. (2026)

Last-iterate convergence

- Guarantees for the regret translate to some form of average convergence
- What can be said about the last iterate of the dynamics?

Common algorithms such as MWU and gradient descent can fail miserably!



Optimism to the rescue

- It turns out that optimism, besides improving the regret, can also ensure last-iterate convergence in some classes of games
- We proved earlier that, in games with nonnegative sum of regrets,

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



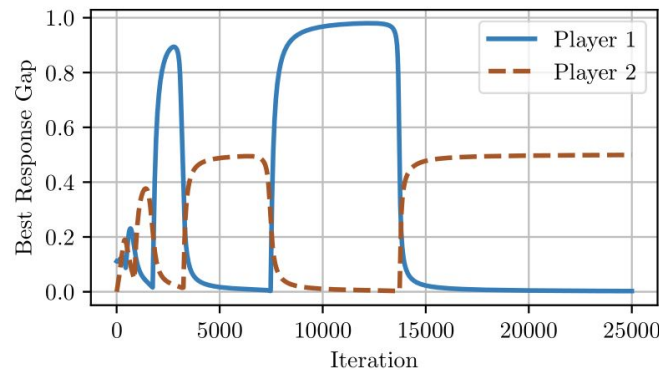
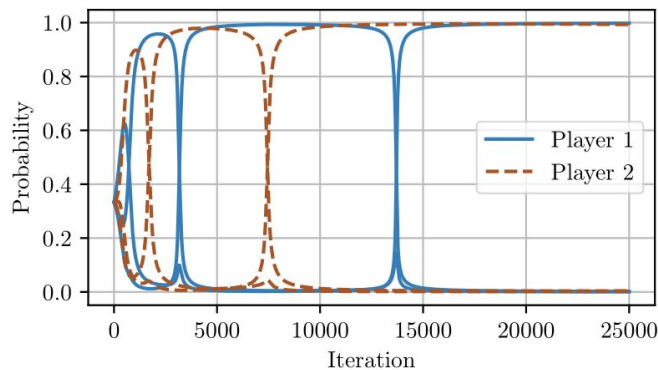
When the regularizer is smooth, small variation implies convergence to Nash equilibria.

Convergence of optimistic learning

Theorem. For any game with nonnegative sum of regrets, after $T = O(1/\epsilon^2)$ rounds, most strategies are $O(\epsilon)$ -Nash equilibria.

- This rate is known to be tight
- Convergence is the ultimate form of predictability, trivializing the problem of online learning
- But what if the dynamics do not converge to Nash equilibria?

Small variation without Nash convergence



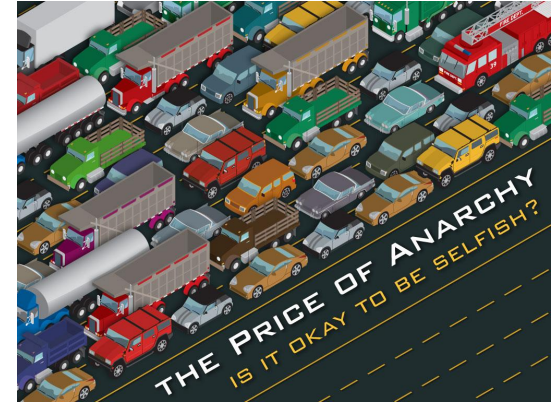
$$\sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}\|_1^2 \leq O(\log T).$$



Small variation doesn't necessarily imply convergence to Nash equilibria

Social welfare and smooth games

- As we have seen, some equilibria are better than others
- Is regret minimization converging to good equilibria?
- As is standard, we measure goodness through social welfare, although there are many other ways to quantify the quality of equilibria
- The framework of **price of anarchy** quantifies the inefficiency of equilibria



From Roughgarden

Smooth games

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

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

- If each player follows its component from the welfare-optimal strategy, the players collectively get some fraction of the optimal welfare
- Many classes of games are known to be smooth (Roughgarden, 2015)
- In some sense, a generalization of zero-sum games

Connection with regret minimization

In any smooth game,

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

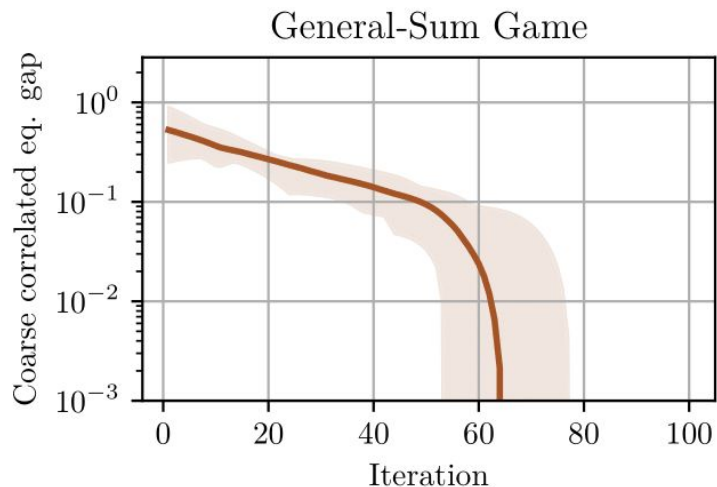
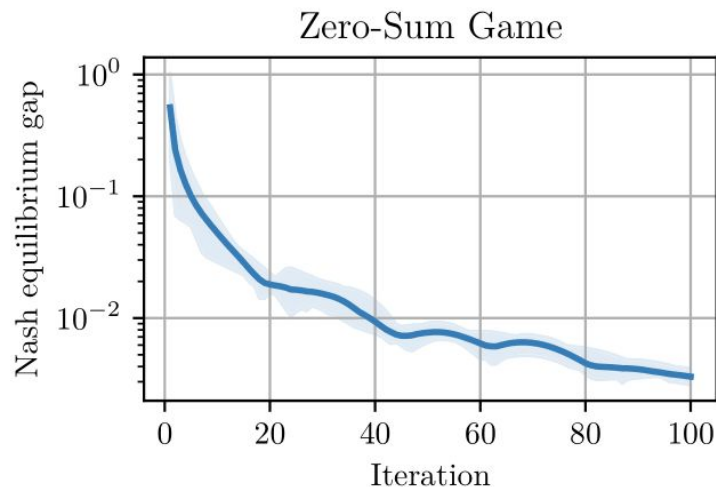
- Convergence to a near-optimal equilibrium is driven by the *sum* of the players' regrets
- The ratio $\lambda/(1+\mu)$ is called **robust price of anarchy**
- What if the regrets are negative?

Optimistic mirror descent in smooth games

Theorem. Optimistic mirror descent

- Either converges to a Nash equilibrium
 - Or the average welfare outperforms the robust price of anarchy
-
- Individually each problem is hard!
 - The further away from Nash equilibria, the larger the improvement in terms of the social welfare

Optimistic mirror descent in two-player games



Optimistic mirror descent in two-player games

