Faster No-Regret Learning Dynamics and Last-Iterate Convergence

15 888 **Computational Game Solving** (Fall 2025)

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

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

→ We need new algorithmic ideas!

Optimistic regret minimization

Optimistic FTRL

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

Optimistic mirror descent

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

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

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

Regret bounded by variation in utilities

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

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)

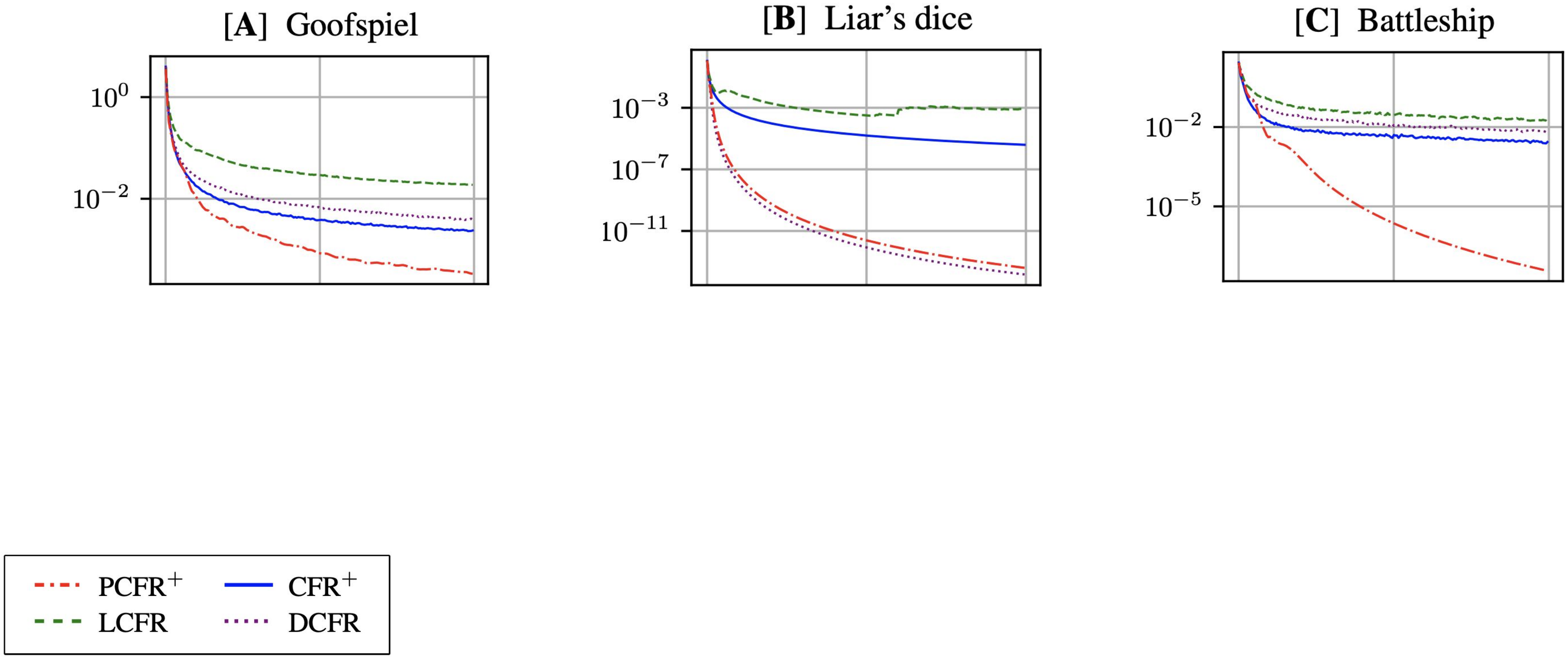
Initialize cumulative regrets $\boldsymbol{r}^{(0)} \coloneqq \boldsymbol{0}$;
for $t = 1, \dots, T$ **do**
    Define $\boldsymbol{\theta}^{(t)} \coloneqq [\boldsymbol{r}^{(t-1)} + \boldsymbol{m}^{(t)} - \langle \boldsymbol{m}^{(t)}, \boldsymbol{x}^{(t-1)} \rangle \boldsymbol{1}]^+$;
    if $\boldsymbol{\theta}^{(t)} = \boldsymbol{0}$ **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 $\boldsymbol{u}^{(t)} \in [0, 1]^{\mathcal{A}}$;
    $\boldsymbol{r}^{(t)} \coloneqq \boldsymbol{r}^{(t-1)} + \boldsymbol{u}^{(t)} - \langle \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle \boldsymbol{1}$;

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

Initialize cumulative regrets $\boldsymbol{r}^{(0)} \coloneqq \boldsymbol{0}$;
for $t = 1, \dots, T$ **do**
    Define $\boldsymbol{\theta}^{(t)} \coloneqq [\boldsymbol{r}^{(t-1)} + \boldsymbol{m}^{(t)} - \langle \boldsymbol{m}^{(t)}, \boldsymbol{x}^{(t-1)} \rangle \boldsymbol{1}]^+$;
    if $\boldsymbol{\theta}^{(t)} = \boldsymbol{0}$ **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 $\boldsymbol{u}^{(t)} \in [0, 1]^{\mathcal{A}}$;
    $\boldsymbol{r}^{(t)} \coloneqq [\boldsymbol{r}^{(t-1)} + \boldsymbol{u}^{(t)} - \langle \boldsymbol{x}^{(t)}, \boldsymbol{u}^{(t)} \rangle \boldsymbol{1}]^+$;

PRM in action

From Farina et al.



Faster rates using stability

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

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

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

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

- 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

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

- 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

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

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

Games with nonnegative sum of regrets

Theorem. For any game with nonnegative sum of regrets,

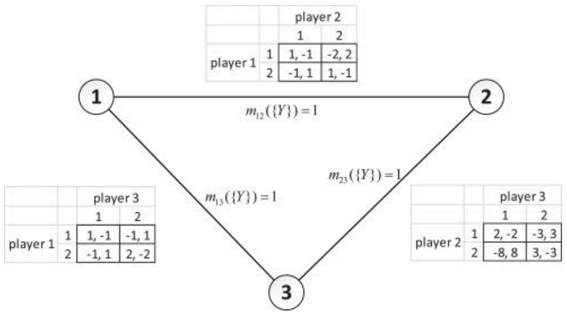
$$\sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \leq O(1).$$

The misprediction error is bounded!

The assumption holds for zero-sum games:

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

Polymatrix zero-sum games



Taken from Deng et al.

- 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

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

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

A reminder of Blum-Mansour

Algorithm 2: Blum-Mansour algorithm for minimizing swap regret

Input: A regret minimizer $\mathfrak{R}_a$ for each action $a \in \mathcal{A}$

$\textsc{NextStrategy}()$:

  for *each action* $a \in \mathcal{A}$ **do**

    $\Delta(\mathcal{A}) \ni \boldsymbol{x}_a^{(t)} \coloneqq \mathfrak{R}_a.\textsc{NextStrategy}()$;

  Set $\mathbf{M}^{(t)} \coloneqq [(\boldsymbol{x}_a^{(t)})_{a \in \mathcal{A}}]$;

  return $\Delta(\mathcal{A}) \ni \boldsymbol{x}^{(t)} = \mathbf{M}^{(t)} \boldsymbol{x}^{(t)}$;

$\textsc{ObserveUtility}(\boldsymbol{u}^{(t)} \in \mathbb{R}^{\mathcal{A}})$:

  for *each action* $a \in \mathcal{A}$ **do**

    Set $\boldsymbol{u}_a^{(t)} \coloneqq \boldsymbol{x}^{(t)}[a] \boldsymbol{u}^{(t)}$;

    $\mathfrak{R}_a.\textsc{ObserveUtility}(\boldsymbol{u}_a^{(t)})$;

RVU bound for Blum-Mansour

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

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

It suffices to prove

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

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**

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

- 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:

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

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

Improved regularizer

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

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

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

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

RVU for swap regret

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

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

Corollary.

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

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

- Improving this to a constant is an open question

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

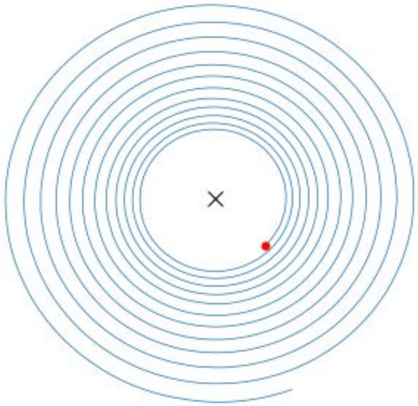
$$\sum_{\tau=1}^{t} \|\boldsymbol{u}^{(\tau)} - \boldsymbol{u}^{(\tau-1)}\|_*^2.$$

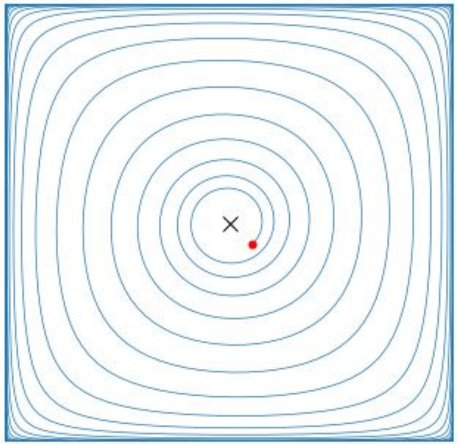
☞ If it gets larger than logarithmic, we can switch to an algorithm tuned for the adversarial regime

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,

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

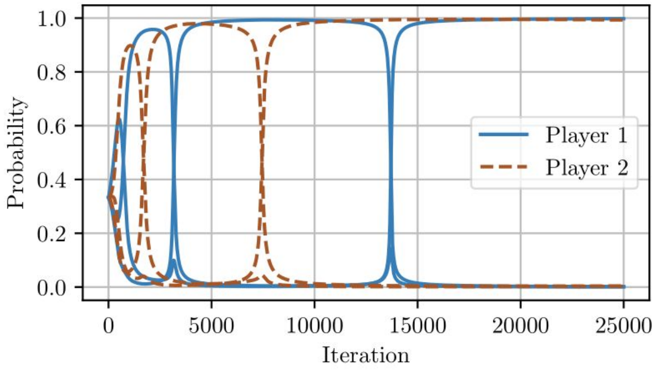
Under some assumptions, small variation implies convergence to Nash equilibria.

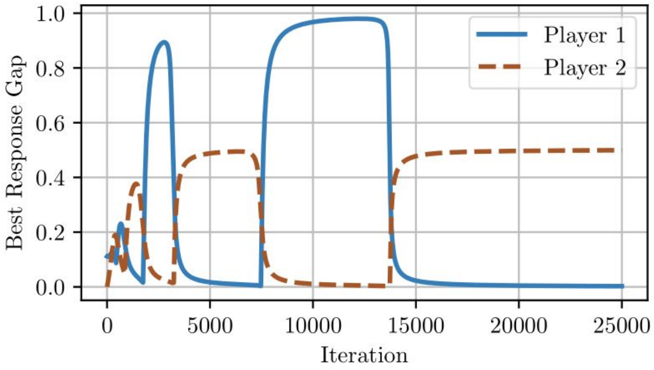
Convergence of optimistic learning

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



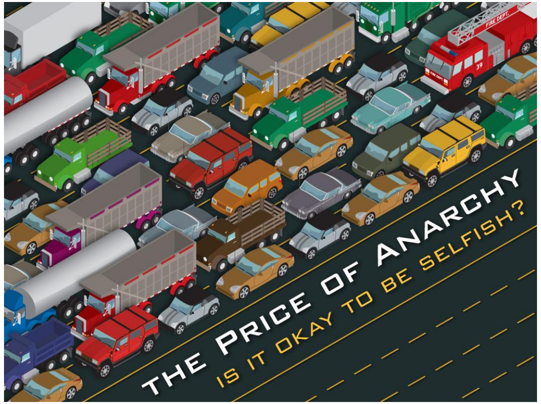


$$\sum_{i=1}^{n} \sum_{t=1}^{T} \|\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)}\|_1^2 \le O(\log T).$$

☞ 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 $(\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n)$ if

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

- 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,

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

- 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?

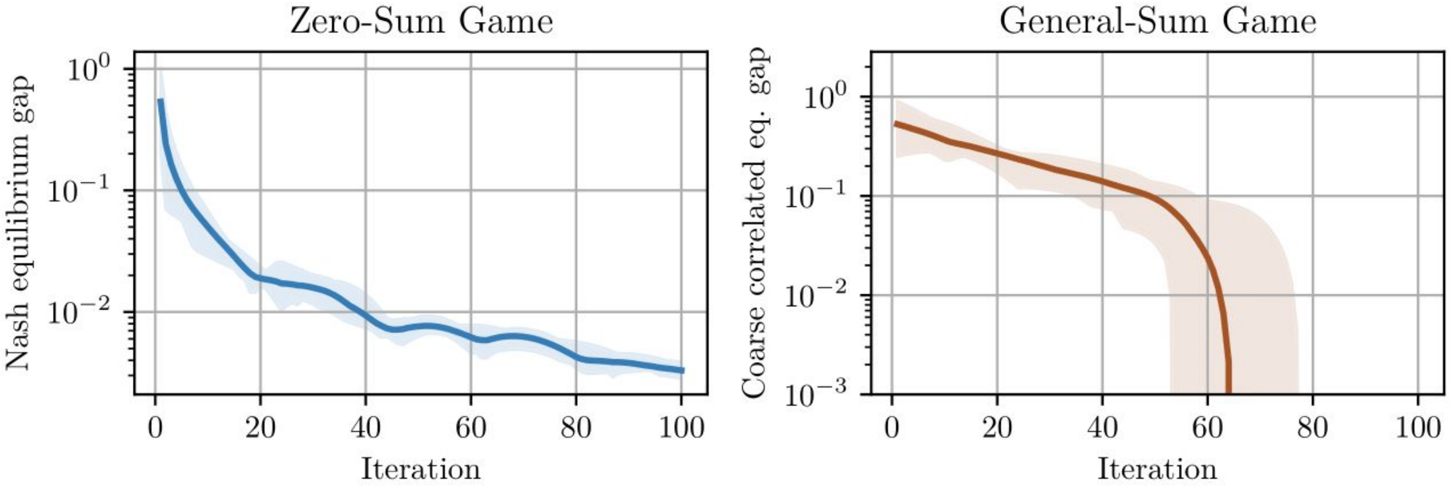
Optimistic mirror descent in smooth games

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