

Lecture 4: Game solving via online learning

3 Sept 2025

CS6208 Fall 2025: Computational Game Theory

Admin Matters

Homework 1 should be out soon (today or tomorrow)

- Make sure to find teammate(s)

Survey on teaching

Tutorials?

Recall from last week...

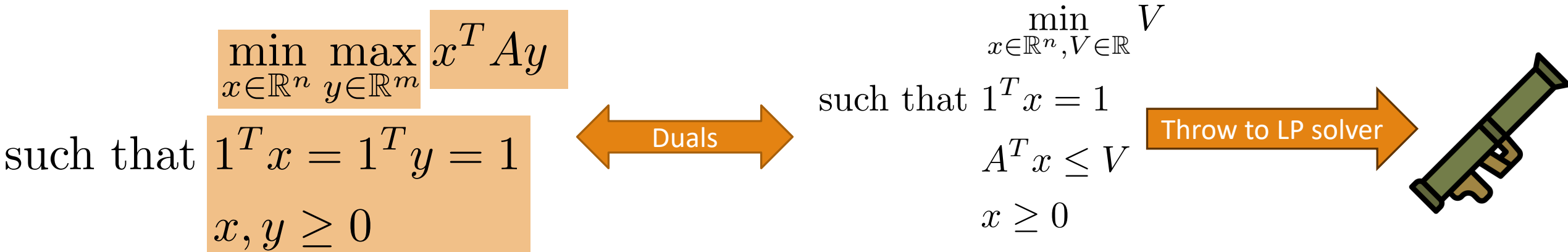
Nash as a minimax problem

The Von Neumann minimax theorem

Solving zero-sum games using linear programming

Unresolved issues

- Poly time, but how fast exactly?
- I cheated by outsourcing the problem to LP solver
- Seems to be overkill? LP and games are closely linked, but LP *solving* seems too general an algorithm? Is there something more intuitive?



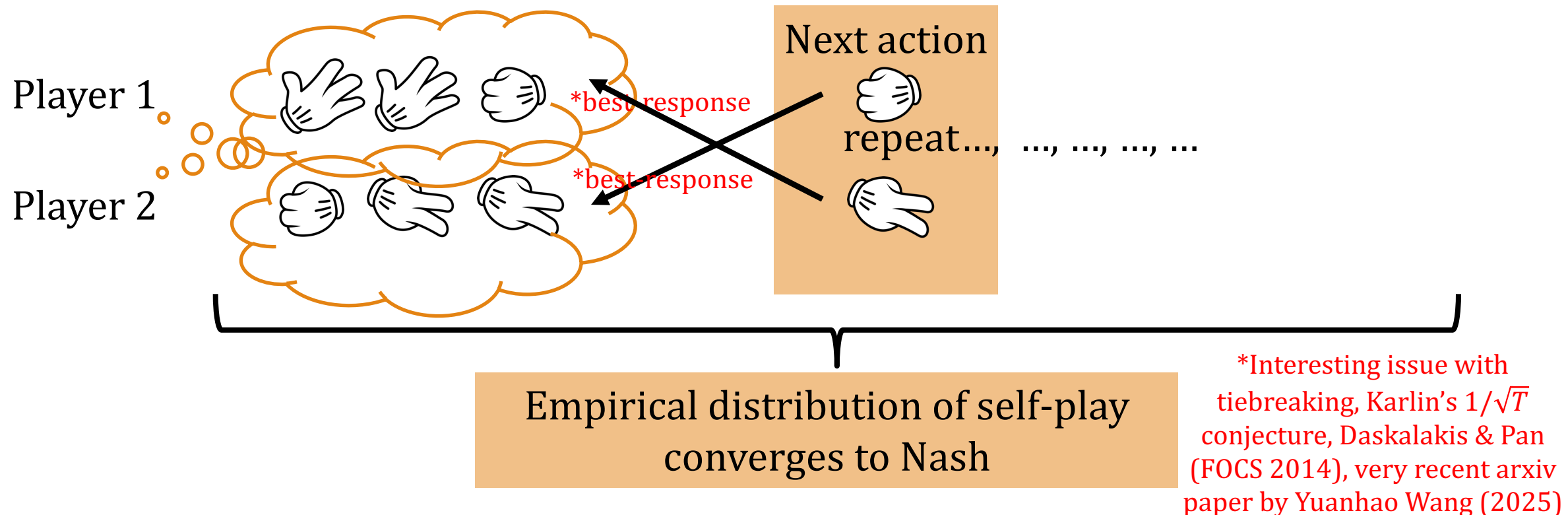
Warmup: Fictitious play

Recall: $A = -B$: Your pain is my pleasure

- Nice properties (e.g., exchangeability, unique game value, polytime solvers)

Fictitious Play (Brown, 1951): *Self play* can lead to Nash

- Players best-respond to the *empirical distribution* of past opponent actions



Quality of Equilibrium

Comparing Different Strategies

Which strategy is “better”?

1. [Rock = 0.4, Paper = 0.3, Scissors = 0.3]
2. [Rock = 0, Paper = 1, Scissors = 0]

Strategy 2 gets on average a payoff of 0.1 when playing against Strategy 1

But...

Strategy 2 “looks” very far from the Nash (i.e., uniform strategy)

Cannot order strategies in terms of head-on performance

Very common mistake

Discussion

Which is the better poker player?

- A: someone who earns a lot of money on average against the general population, but loses slightly against the strongest opponents
- B: someone who earns less overall, but beats everyone, even top professionals

Philosophical question

- Do you really agree that B is a better player?
- Will you say that not just to me, but to your peers?

Saddle Point Residual

If x_0, y_0 are a **pair** of strategies, the saddle point residual (or gap) is

$$\underbrace{\max_y x_0^T A y - x_0^T A y_0}_{\text{Extra amount that max player could have obtained if it best-responded "in hindsight" AKA regret}} + \underbrace{x_0^T A y_0 - \min_x x^T A y_0}_{\text{Extra amount that min player could have obtained if it best-responded "in hindsight" AKA regret}}$$

$$= \max_y x_0^T A y - \min_x x^T A y_0$$

Intuition: if both players don't regret playing the best response in hindsight, what do we have?

Approximate Equilibrium

(x_0, y_0) is a an ϵ -approximate Nash (ϵ - Nash) if expected utility from unilaterally deviating does not increase by more than ϵ

$$\min_{x \in \Delta_n} x^T A y_0 \geq x_0^T A y_0 - \epsilon \quad \text{AND} \quad \max_{y \in \Delta_m} x_0^T A y \leq x_0^T A y_0 + \epsilon$$

- Sanity Check: A NE itself is a 0-approximate NE
- There is another definition (well supported approximated equilibrium) based on what actions are allowed in the support

extends to general-sum NE also

Saddle point residual = $\epsilon \rightarrow \epsilon$ -approximate NE

$$\min_{x \in \Delta_n} x_0^T A y_0 - x^T A y_0 = \epsilon_1 \quad \max_{y \in \Delta_m} x_0^T A y - x_0^T A y_0 = \epsilon_2$$

$$\max_{y \in \Delta_m} x_0^T A y - \min_{x \in \Delta_n} x^T A y_0 = \epsilon_1 + \epsilon_2 \geq \max(\epsilon_1, \epsilon_2)$$

Let's minimize the saddle-point residual instead!

No-Regret Learning

“In the end... we only regret the chances we didn't take, the relationships we were afraid to have, and the decisions we waited too long to make.”

-Lewis Carroll

What we will be working towards

Fictitious play

- Simulate 2 players playing against each other
- Best respond against the average history of the other player

We will do something more general here, by setting *no-regret learners* against each other.

- No-regret learning will imply that the iterates will in some sense converge to Nash (for 2-player zero-sum games)
- Fictitious play is known as Follow-the-Leader, which is *not* no-regret

No-regret learning is an area of interest even outside game theory

- Exist lots of cross fertilization, e.g., sequential decision making,
 - Trust-region policy optimization in RL is no-regret!
- No-regret learning remains state-of-the-art for equilibrium finding
- Study of dynamics between no-regret learners is an active and exciting area of research (drop me an email for discussion)

The basic setting

$$\mathcal{X} \in \Delta_n \subseteq \mathbb{R}^n,$$

At each iteration t

- Learner chooses $x^{(t)} \in \Delta_n$
- Observe loss/reward vector $g^{(t)} \in [0,1]^n$
- Reward at time $t = \langle x^{(t)}, g^{(t)} \rangle$

Important: choose $x^{(t)}$
before observing $g^{(t)}$

$$\langle a, b \rangle = a^T b$$

The $g^{(t)}$ are chosen by an **adversary**

- It knows your algorithm, can “simulate” your choice of $x^{(t)}$
- **Caution:** another formulation using for random *actions* (oblivious adversary)

What is a smart way of choosing $x^{(t)}$?

What is a good metric?

- Option 1: compare to best sequence of $\{x^{(t)}\}$ in hindsight
- Option 2: compare to best single x in hindsight

What baseline to compare to?

Option 1: best sequence in hindsight

$$\underbrace{\sum_{t=1}^T \langle x^{(t)}, g^{(t)} \rangle}_{\text{What we got}} - \underbrace{\sum_{t=1}^T \min_x \langle x, g^{(t)} \rangle}_{\text{What we could have got}}$$

- Remember adversary knows what $x^{(t)}$ we can choose!
- Let $d = n$. If $x_1^{(t)} \geq 0.5$, adversary sets $g^{(t)} = (0, 1)$ else $g^{(t)} = (1, 0)$
 - Loss we get is always less than half!

$$\sum_{t=1}^T \langle x^{(t)}, g^{(t)} \rangle - \sum_{t=1}^T \min_x \langle x, g^{(t)} \rangle = \sum_{t=1}^T \underbrace{\langle x^{(t)}, g^{(t)} \rangle - \min_x \langle x, g^{(t)} \rangle}_{\geq 0.5}$$

- We perform, compared to the best sequence in hindsight, $0.5T$ as well
 - Want this to be sublinear in T , if so, then as $T \rightarrow \infty$ we perform better on average

Let's build some
intuition (investing)

Is this a good fund manager?

Suppose you have a pot of money you want to invest

- Consult a manager, park some money with him
- 1 year later, he comes back to you with +15% returns

Is this a good or bad performance? Should you fire your fund manager?

- 30-40 years ago, 15% is probably bad, though 15% pa is decent enough now
- What do people normally say? Counterfactuals

“I could have done XYZ instead and outperformed my fund manager”

What are some reasonable “XYZs”



Temasek annual shareholder return of 1.6%.

Other

I think it is pretty bad right? Temasek just need to park the funds in S&P500 to get 15-20% return.

<https://www.channelnewsasia.com/singapore/temasek-review-portfolio-value-us-india-china-investment-4465621>



Short-Panic-2820 • 2mo ago

Ya lor why dosent Temasek just open their ibkr app and put in a market order for \$389b worth of vwra? Are they stupid??

⊖ ↑ 401 ↓ 💬 Reply 🏆 Award ➦ Share ...



likpopper • 2mo ago

S and p did 28% at the same time range

↑ 18 ↓ 💬 Reply 🏆 Award ➦ Share ...

For reference, **S&P**500 had 7.7% (inflation-adjusted) annual return over the past 20 year period. Mind you this is an ETF, meaning it is not actively managed, while Temasek is an actively-managed fund which means more work done on research, compliance etc ie more money to pay the people managing the investments, only for them to underperform the broader economy.

Important: I am not giving investment advice!
There are many reasons for why one won't want to adopt regret minimization for their portfolio...

Comparing to best sequence in hindsight is too harsh

But people often compare against best **fixed strategy in hindsight**

~~$$\sum_{t=1}^T \langle x^{(t)}, g^{(t)} \rangle - \sum_{t=1}^T \min_x \langle x, g^{(t)} \rangle$$~~

$$\sum_{t=1}^T \langle x^{(t)}, g^{(t)} \rangle - \min_x \sum_{t=1}^T \langle x, g^{(t)} \rangle$$

Okay, back to student-
poverty land

Regret minimization needs randomness

$$R_T = \sum_{t=1}^T \langle x^{(t)}, g^{(t)} \rangle - \min_x \sum_{t=1}^T \langle x, g^{(t)} \rangle$$

Intuition: play action that has done well in the past more, e.g., if S&P did well in the past, then put more of our portfolio in S&P

- Consider an adversary's point of view
 - If it still chooses S&P to do well, then our average regret (we are now doing well)
 - If it tries to punish the learner by making S&P perform bad, then yes, we still do badly, but S&P as a best-in-hindsight option becomes worse

Follow-the-leader algorithm (what we did for fictitious play)

- Choose the action with the lowest loss always

$$\arg \min_{x \in \{e_1, \dots, e_n\}} \sum_{t=1}^T \langle x, g^{(t)} \rangle$$

- Doesn't work. No deterministic strategy works. Adversary chooses loss = 1 at chosen coordinate and 0 everywhere else. Expected loss = T, best in hindsight at most T/n (pigeonhole principle)

Hedge (aka multiplicative weights)

We are going to maintain a *weight* for each action $w_i^{(t)} \in \mathbb{R}$

Play according to weighted average at each timestep

$$x^{(t)} = \frac{w^{(t)}}{\sum_i w_i^{(t)}} \quad \leftarrow \text{Vector of length } n$$

Update weights according to

$$w_i^{(t+1)} = w_i^{(t)} \exp(-\eta \cdot g_i^{(t)}) \quad \leftarrow \text{Learning rate } > 0$$

Sanity check: if $g_i^{(t)}$ is high (i.e., big losses), then $\exp(-\eta \cdot g_i^{(t)})$ is small \rightarrow down-weight that action for iterate $t + 1$ (and beyond)

Initialize $w^{(1)}$ as the all-ones vector

Implementation detail: store $\sum_{\tau} g^{(\tau)}$ instead, and take softmax when computing $x^{(t)}$

Regret bound for Hedge

We will prove $R_T \leq \frac{\log(n)}{\eta} + \frac{\eta T}{2}$

Setting $\eta = 1/\sqrt{T}$ gives $\underbrace{O(\log(n) \cdot \sqrt{T})}_{\text{Optimal in } n \text{ and } T}$ regret

What to do when T is unknown
beforehand? Doubling trick.
Time-varying learning rate

Shows that no-regret minimizers exist!

In fact, the existence of no-regret minimizers can be used to constructively prove the minimax theorem

- at least for bilinear functions, though can be generalized

Hedge is also a variant of

- *Follow-the-regularized-leader* (FTRL)
- *Online mirror descent* (OMD) with appropriately chosen DGFs

Proof of regret bounds

Let $Z_t = \sum_{j=1}^n w_j^{(t)}$, i.e., the total weight at time t

$$Z_{t+1} = \sum_{j=1}^n w_j^{t+1}$$

$$= \sum_{j=1}^n w_j^t \cdot \exp(-\eta g_j^{(t)}) \quad \text{Definition of } w^{(t)}$$

$$= Z_t \cdot \sum_{j=1}^n \frac{w_j^t}{Z_t} \cdot \exp(-\eta g_j^{(t)}) \quad Z_t > 0$$

$$= Z_t \cdot \sum_{j=1}^n x^{(t)} \cdot \exp(-\eta g_j^{(t)}) \quad \text{Definition of } x^{(t)}$$

$$\leq Z_t \cdot \sum_{j=1}^n x^{(t)} \cdot (1 - \eta g_j^{(t)} + \frac{\eta^2}{2} (g_j^{(t)})^2) \quad \text{2nd order expansion}$$

$$= Z_t \cdot (1 - \eta \langle x^{(t)}, g^{(t)} \rangle + \frac{\eta^2}{2} \langle x^{(t)}, (g^{(t)})^2 \rangle)$$

$$\leq Z_t \cdot \exp(-\eta \langle x^{(t)}, g^{(t)} \rangle + \frac{\eta^2}{2} \langle x^{(t)}, (g^{(t)})^2 \rangle) \quad g \in [0,1]$$

$$Z_{t+1} \leq n \cdot \prod_{\tau=1}^t \exp \left(-\eta \langle x^{(\tau)}, g^{(\tau)} \rangle + \frac{\eta^2}{2} \langle x^{(\tau)}, (g^{(\tau)})^2 \rangle \right) \quad \text{telescoping}$$

$$= n \cdot \exp \left(-\eta \sum_{\tau=1}^t \langle x^{(\tau)}, g^{(\tau)} \rangle + \frac{\eta^2}{2} \sum_{\tau=1}^t \langle x^{(\tau)}, (g^{(\tau)})^2 \rangle \right)$$

Suppose i^* is the best action in hindsight, then.

$$\exp \left(-\eta \underbrace{\sum_{\tau=1}^t g_{i^*}^{(\tau)}}_{\text{What we could have gotten}} \right) = w_{i^*}^{(t+1)} \leq Z_{t+1}$$

What we could have gotten

$$\leq n \cdot \exp \left(-\eta \underbrace{\sum_{\tau=1}^t \langle x^{(\tau)}, g^{(\tau)} \rangle}_{\text{Loss encountered}} + \frac{\eta^2}{2} \sum_{\tau=1}^t \langle x^{(\tau)}, (g^{(\tau)})^2 \rangle \right)$$

Loss encountered

Take log, rearrange

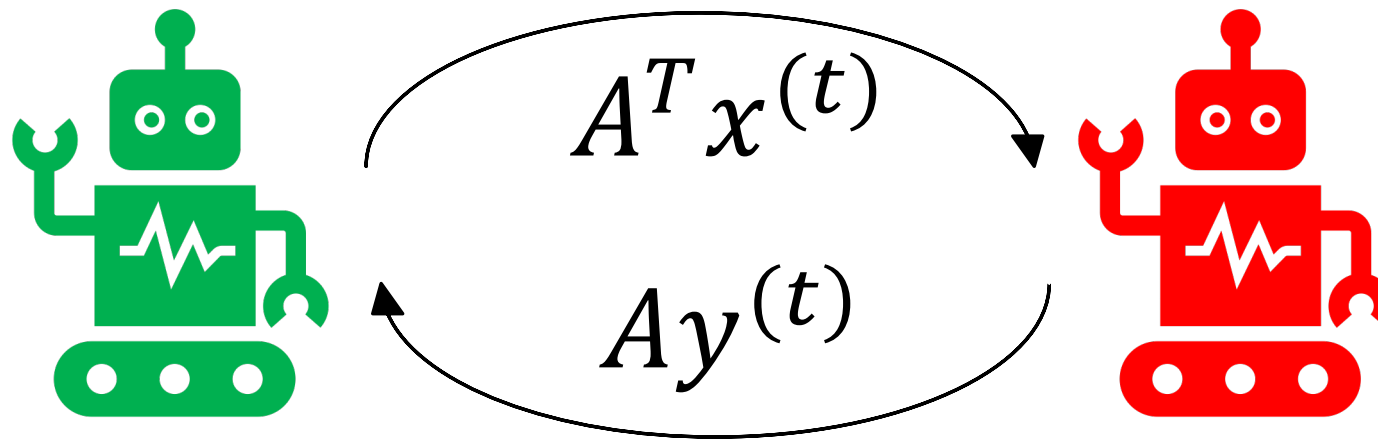
$$R_t \leq \frac{\log(n)}{\eta} + \frac{\eta}{2} \sum_{\tau=1}^t \langle x^{(t)}, g^{(t)} \rangle \leq \frac{\log(n)}{\eta} + \frac{\eta t}{2}$$

Going from no-regret learning to NE

I swear we are going somewhere with this

A General Framework for Self-Play

$\text{NEXTSTRATEGY}()$
 $\text{OBSERVEUTILITY}(\ell^{(t)})$ } Loop



Average strategies converge to Nash (saddle point residual drops to 0)

Self-play with no-regret algorithms

Suppose we have a low total regret, i.e.,

$$\max_y \sum_{\tau=1}^t \langle x^{(\tau)}, Ay \rangle - \langle x^{(\tau)}, Ay^{(\tau)} \rangle \leq R_1$$

$$\min_x \sum_{\tau=1}^t \langle x^{(\tau)}, Ay^{(\tau)} \rangle - \langle x, Ay^{(\tau)} \rangle \leq R_2$$

Summing the inequalities, we get

$$\max_y \sum_{\tau=1}^t \langle x^{(\tau)}, Ay \rangle - \min_x \sum_{\tau=1}^t \langle x, Ay^{(\tau)} \rangle \leq R_1 + R_2$$

Averaging

$$\max_y \left\langle \frac{\sum_{\tau=1}^t x^{(\tau)}}{t}, Ay \right\rangle - \min_x \left\langle x, A \frac{\sum_{\tau=1}^t y^{(\tau)}}{t} \right\rangle \leq \frac{R_1 + R_2}{t}$$

Sublinear in T !

Average strategies have low saddle point gap \rightarrow approximate NE!

Learning NE from self-play

Instantiate self-play using any pair of regret minimizers!

For Hedge, saddle-point residual drops at rate

$$\mathcal{O}(\log(\max(n, m)) \cdot \frac{1}{\sqrt{T}})$$

In practice, we use another regret minimizer called *Regret Matching*

- Store regret vector (over all alternative actions)
- Choose action **in proportion to regret** (ignore actions with -ve regret)
- RM⁺ permanently adjusts negative regrets to 0 at each timestep

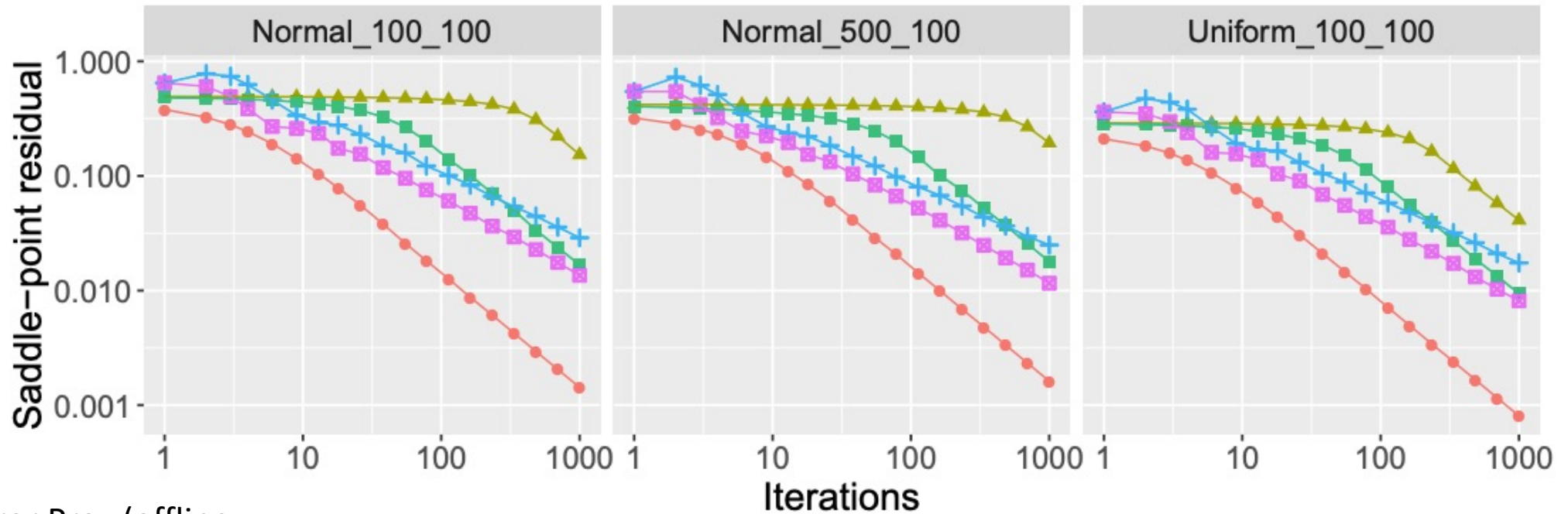
RM (and RM⁺) performs much better in practice, but has rate

$$\mathcal{O}(\sqrt{\max(n, m)} \cdot \frac{1}{\sqrt{T}})$$

These rates are obtained using our basic analysis of self-play, which may not be tight

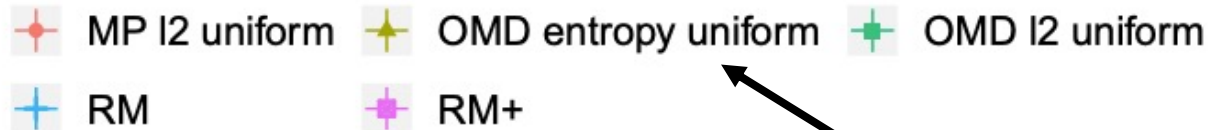
- RM is not optimal in n for simplexes

Example convergence rates



Mirror Prox (offline method with $1/T$ convergence)

Algorithm

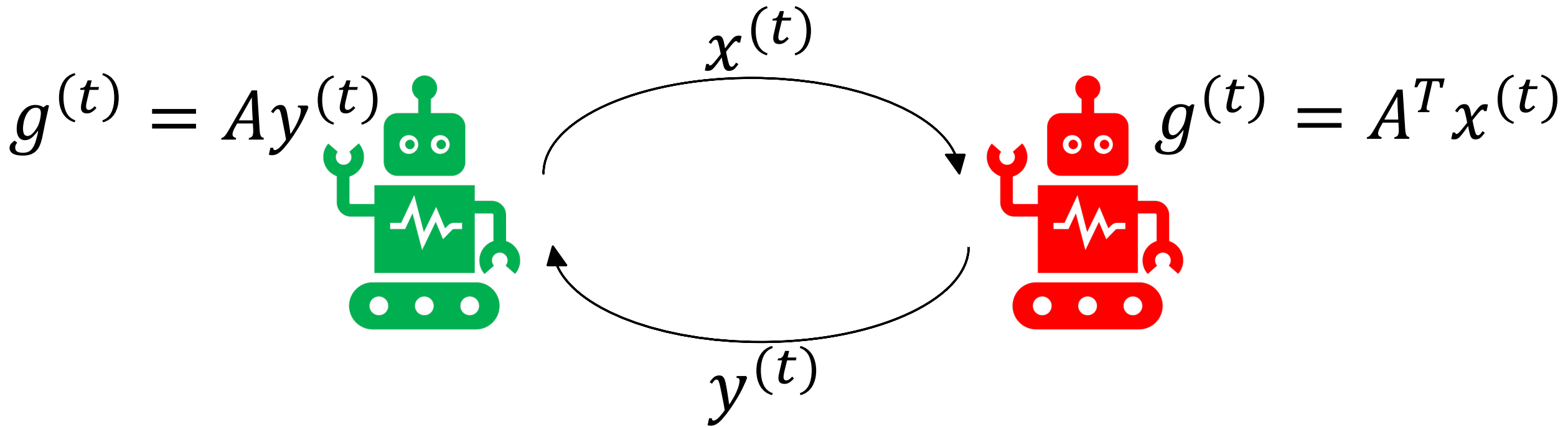


MW/Hedge

Source: http://www.columbia.edu/~ck2945/files/main_ai_games_markets.pdf

Review: self-play [simultaneous]

NEXTSTRATEGY()
OBSERVELOSS($g^{(t)}$) } Loop

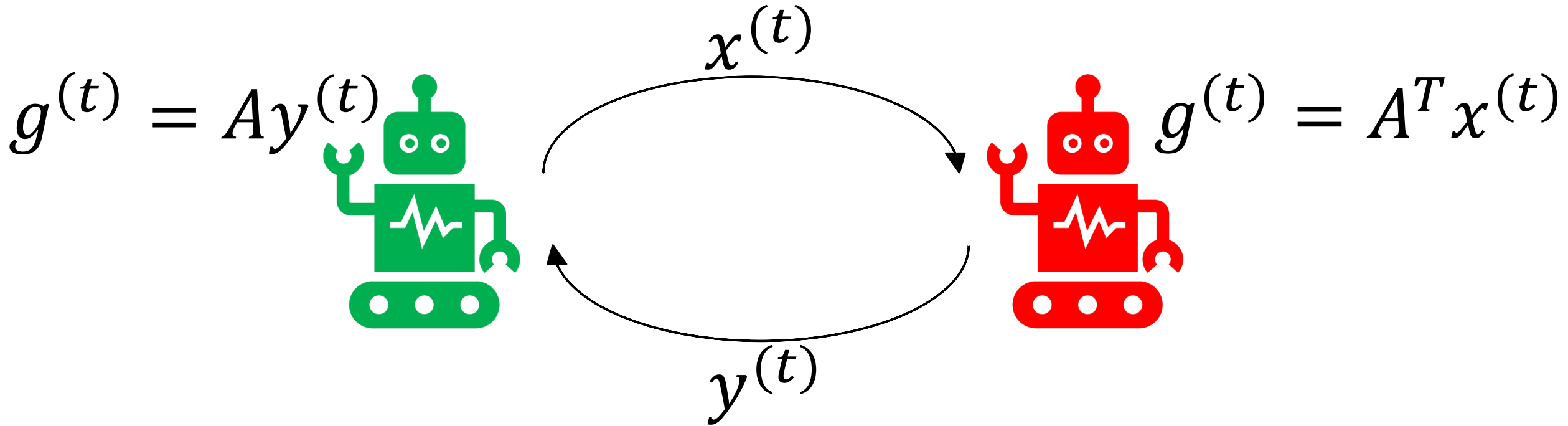


Average strategies converge to Nash (saddle point residual drops to 0)

Review: self-play [alternating]

Under certain mild conditions,
will provably converge faster
than simultaneous play, in
practice, around 3-10x

NEXTSTRATEGY()
OBSERVELOSS($g^{(t)}$) } Loop but alternate
between players



Average strategies converge to Nash (saddle point residual drops to 0)

Review: Rate of convergence

Sum of regrets is
sublinear in t

$$\text{Recall: Saddle point residual} \leq \frac{R_1 + R_2}{t}$$

Instantiate self-play using **any pair** of regret minimizers!

For Hedge, saddle-point residual drops at rate

$$\mathcal{O}(\log(\max(n, m)) \cdot \frac{1}{\sqrt{T}})$$

$$\text{Recall: Saddle point residual} \leq \epsilon \implies \epsilon\text{-NE}$$

Review: regret minimization

Expert 1

Expert 2

Expert 3

Player 🧐

$$x_1^{(1)} = 0.25$$

$$x_2^{(1)} = 0.5$$

$$x_3^{(1)} = 0.25$$

Adversary 😈

$$g_1^{(1)} = 1$$

$$g_2^{(1)} = 0.5$$

$$g_3^{(1)} = 0.8$$

$$\text{Loss incurred at time 1} = \langle x^{(1)}, g^{(1)} \rangle = 0.7$$

Best **expert-so-far** in hindsight = 0.5

$$\text{Total regret} = 0.7 - 0.5 = 0.2$$

Player 🧐

$$x_1^{(2)} = 0.1$$

$$x_2^{(2)} = 0.7$$

$$x_3^{(2)} = 0.2$$

Adversary 😈

$$g_1^{(2)} = 0.1$$

$$g_2^{(2)} = 0.8$$

$$g_3^{(2)} = 0.2$$

$$\text{Loss incurred at time 2} = \langle x^{(2)}, g^{(2)} \rangle = 0.61$$

Best **expert-so-far** in hindsight = 1.0

$$\text{Total regret} = 0.61 + 0.7 - 1.0 = 0.31$$

+time



Review: regret minimization

Expert 1

Expert 2

Expert 3

Player 🧐

$$x_1^{(1)} = 0.5$$

$$x_2^{(1)} = 0.5$$

$$x_3^{(1)} = 0.0$$

Adversary 😈

$$g_1^{(1)} = 1$$

$$g_2^{(1)} = 0.1$$

$$g_3^{(1)} = 0.0$$

Loss incurred at time 1 = $\langle x^{(1)}, g^{(1)} \rangle = ?$

Best expert-so-far in hindsight = ?

Total regret = ?

Player 🧐

$$x_1^{(2)} = 0.1$$

$$x_2^{(2)} = 0.3$$

$$x_3^{(2)} = 0.6$$

Adversary 😈

$$g_1^{(2)} = 0.1$$

$$g_2^{(2)} = 0.2$$

$$g_3^{(2)} = 1.0$$

Loss incurred at time 2 = $\langle x^{(2)}, g^{(2)} \rangle = ?$

Best expert-so-far in hindsight = ?

Total regret = ?

+time



Quiz

Recall that we want total regret to be sublinear in time.








Can regret (by our definition) be ever be negative?

Quiz

If we allowed the player to cheat by observing the adversary's choice of $\ell^{(t)}$ before choosing $x^{(t)}$, does the resultant sequence of x 's achieve sublinear total regret?

Player 🧐	$x_1^{(1)} = 0.5$	$x_2^{(1)} = 0.5$	$x_3^{(1)} = 0.0$
Adversary 😈	$g_1^{(1)} = 1$	$g_2^{(1)} = 0.1$	$g_3^{(1)} = 0.0$
Adversary 😈	$g_1^{(1)} = 1$	$g_2^{(1)} = 0.1$	$g_3^{(1)} = 0.0$
Player 🧐	$x_1^{(1)} = 0.0$	$x_2^{(1)} = 0.0$	$x_3^{(1)} = 1.0$

Review: Hedge ($\eta = 1$)

	Expert 1	Expert 2	Expert 3	
Weights 	1	1	1	
Player 	$x_1^{(1)} = 1/3$	$x_2^{(1)} = 1/3$	$x_3^{(1)} = 1/3$	
Adversary 	$g_1^{(1)} = 1$	$g_2^{(1)} = 0.5$	$g_3^{(1)} = 0.8$	
<hr/>				
Weights 	$\exp(-\eta \cdot 1) \cdot 1$	$\exp(-\eta \cdot 0.5) \cdot 1$	$\exp(-\eta \cdot 0.8) \cdot 1$	
Player 	$x_1^{(2)} \propto \exp(-\eta)$	$x_2^{(2)} \propto \exp(-\eta \cdot 0.5)$	$x_3^{(2)} \propto \exp(-\eta \cdot 0.8)$	
Adversary 	$g_1^{(2)} = 0.1$	$g_2^{(2)} = 0.8$	$g_3^{(2)} = 0.2$	
<hr/>				
Weights 	$\exp(-\eta \cdot 1.1) \cdot 1$	$\exp(-\eta \cdot 1.3) \cdot 1$	$\exp(-\eta \cdot 1.0) \cdot 1$	

+time

Self-play in general-sum games?

What if we were to run self-play in general-sum games?

- Nothing is stopping us from doing it
- Both players will still have sublinear regret \rightarrow not incentivized to deviate
- Isn't this sound like Nash? Does that mean that we get NE from self-play?

Ans: not quite

- Player strategies may end up *correlated* (this is true even in 0-sum games)

Average of *joint strategies* converges to a coarse-correlated equilibrium (CCE)

- CCE is a superset of Nash
- For zero-sum games they coincide (up to payoff equivalence)!

Regret Matching (Plus)

Another regret minimizer on the simplex

Review: Regret Matching

Can be derived using Blackwell Approachability (skipped)

Maintain at timestep t , $r_i^{(t)}$, the regret associated to action i

- “How much regret I have from doing what I did instead of action i ”

$$r_i^{(t)} = \sum_{\tau=1}^t \langle x^{(\tau)}, g^{(\tau)} \rangle - \sum_{\tau=1}^t g_i^{(\tau)}$$









$$r^{(t)} = \sum_{\tau=1}^t \langle x^{(\tau)}, g^{(\tau)} \rangle \mathbf{1} - \sum_{\tau=1}^t g^{(\tau)} \text{ *in vector form}$$

At time $t + 1$, play

$$x_i^{(t+1)} = \frac{\max(r_i^{(t)}, 0)}{\sum_j \max(r_j^{(t)}, 0)} \quad \left. \vphantom{\frac{\max(r_i^{(t)}, 0)}{\sum_j \max(r_j^{(t)}, 0)}} \right\} \begin{array}{l} \text{Threshold at 0, then} \\ \text{play proportionately} \end{array}$$

If all $r^{(t)} \leq 0$, play uniformly

Review: RM

	Expert 1	Expert 2	Expert 3	
Regrets 	0	0	0	
Player 	$x_1^{(1)}=1/3$	$x_2^{(1)}=1/3$	$x_3^{(1)}=1/3$	+time ↓
Adversary 	$g_1^{(1)}=1$	$g_2^{(1)}=0.5$	$g_3^{(1)}=0.8$	
	Loss incurred at time 1 = 2.3/3			
Regrets 	$r_1^{(1)}=2.3/3-1=-0.233$	$r_2^{(1)}=2.3/3-0.5=0.267$	$r_3^{(1)}=2.3/3-0.8=-0.033$	
Player 	$x_1^{(2)} = 0$	$x_2^{(2)} \propto 0.267 = 1$	$x_3^{(2)} = 0$	
Adversary 	$g_1^{(2)}=0.1$	$g_2^{(2)}=0.8$	$g_3^{(2)}=0.2$	
	Loss incurred at time 2 = 0.8, total loss = 1.67			
Regrets 	$r_1^{(2)}=1.67-1.1=0.57$	$r_2^{(2)}=1.67-1.3=0.37$	$r_3^{(2)}=1.67-1.0=0.67$	
Player 	$x_1^{(3)} \propto 0.57$	$x_2^{(3)} \propto 0.37$	$x_3^{(3)} \propto 0.67$	

Why another regret minimizer?

Isn't Hedge already optimal?

Hedge (technically) depends on a learning rate









- Depends on horizon, can be set carefully, but quite annoying
- Another way is to decay the learning rate
- RM is free of learning rate

Theory vs. practice

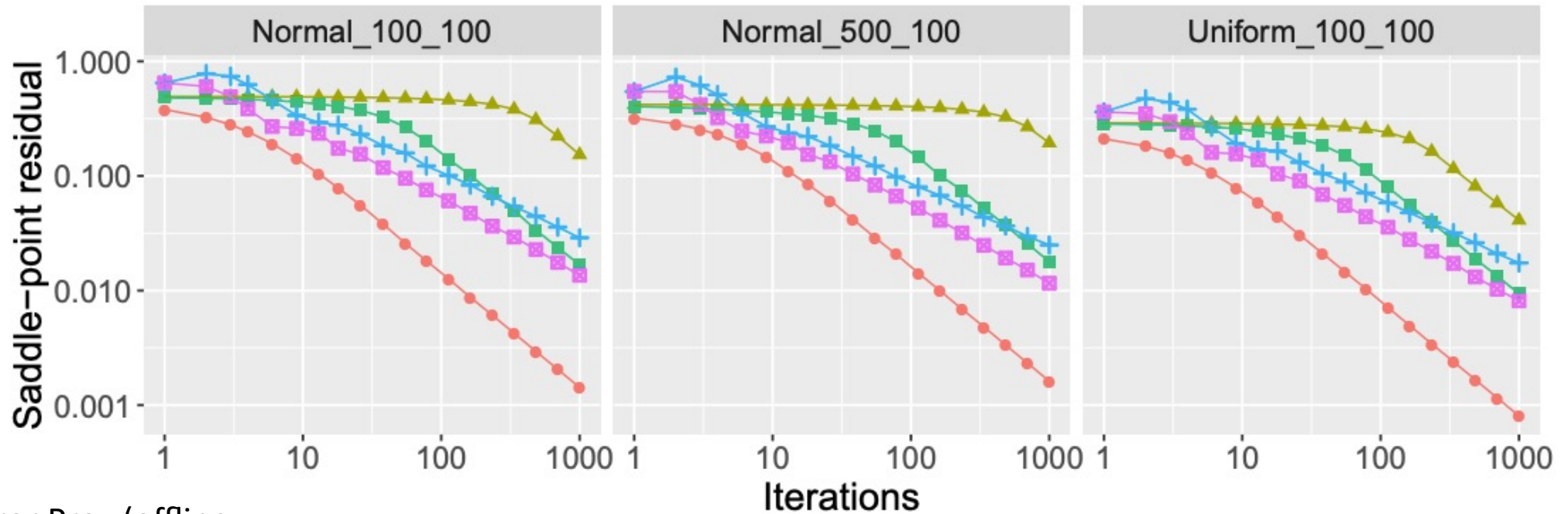
- RM works well in practice
- RM is easy to code, only requires a memory of size n
 - So is Hedge technically...

RM+

Same as RM, but when we threshold regrets at 0, we do it **permanently**

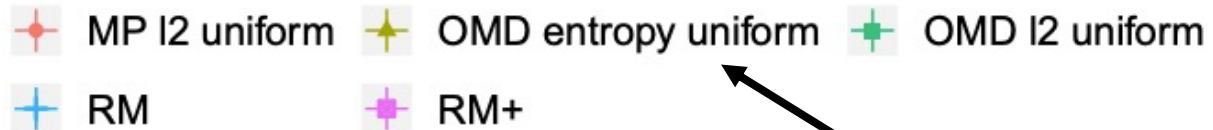
Regrets 	0	0	0
Player 	$x_1^{(1)} = 1/3$	$x_2^{(1)} = 1/3$	$x_3^{(1)} = 1/3$
Adversary 	$g_1^{(1)} = 1$	$g_2^{(1)} = 0.5$	$g_3^{(1)} = 0.8$
	Loss incurred at time 1 = $2.3/3$		
Regrets 	$r_1^{(1)} = 2.3/3 - 1 = -0.233 \rightarrow 0$	$r_2^{(1)} = 2.3/3 - 0.5 = 0.267$	$r_3^{(1)} = 2.3/3 - 0.8 = -0.033 \rightarrow 0$
Player 	$x_1^{(2)} = 0$	$x_2^{(2)} \propto 0.267 = 1$	$x_3^{(2)} = 0$
Adversary 	$g_1^{(2)} = 0.1$	$g_2^{(2)} = 0.8$	$g_3^{(2)} = 0.2$
	Loss incurred at time 2 = 0.8		
Regrets 	$r_1^{(2)} = 0 + 0.8 - 0.1 = 0.7$	$r_2^{(2)} = 0.267 + 0.8 - 0.8 = 0.267$	$r_3^{(2)} = 0 + 0.8 - 0.2 = 0.6$
Player 	$x_1^{(3)} \propto 0.7$	$x_2^{(3)} \propto 0.267$	$x_3^{(3)} \propto 0.6$

Example convergence rates



Mirror Prox (offline method with $1/T$ convergence)

Algorithm



MW/Hedge

Source: http://www.columbia.edu/~ck2945/files/main_ai_games_markets.pdf

Blackwell Approachability (optional)

We are going to construct RM, a more **practical** regret minimizer

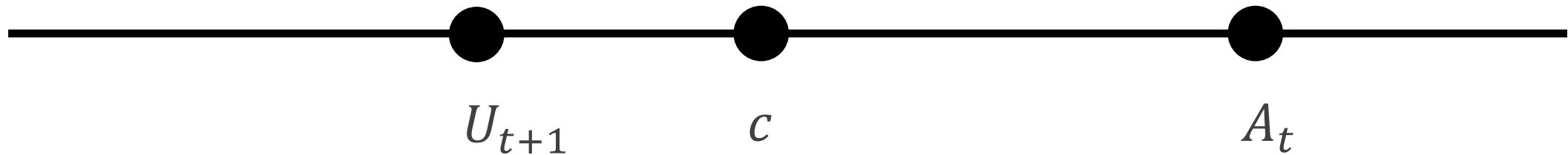
Approachability in Scalars

Sequence of **bounded** scalars $\{U_t\}$, $U_t \in \mathbb{R}$

Let average be $A_T = \frac{1}{T} \sum_{t=1}^T U_t$

Let $c \in \mathbb{R}$ be a target.

Assume $\{U_t\}$ is constrained such that $(U_{T+1} - c)(A_T - c) \leq 0$



Then $\lim_{T \rightarrow \infty} A_T = c$

Intuition: being on the “opposite” side gives enough “power” to reach c , boundedness of U ensures no oscillations.

Approachability in Vectors

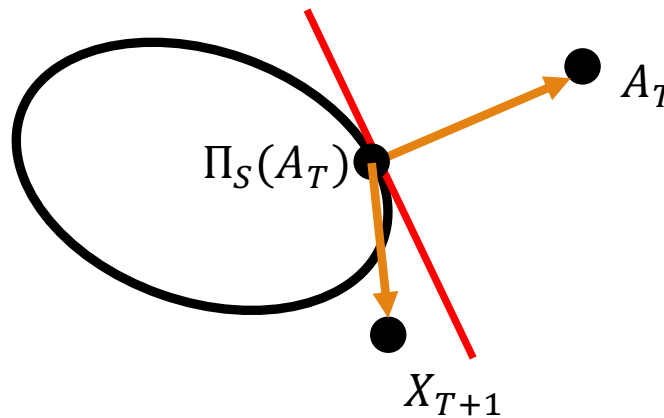
Sequence of **bounded** vectors $\{U_t\}$, $U_t \in \mathbb{R}^K$

Let average be $A_T = \frac{1}{T} \sum_{t=1}^T U_t$

Let $S \in \mathbb{R}$ be a **convex target set**.

- Let $\Pi_S(A_t)$ be the closest point (projection) of A_t onto S

Assume $\{U_t\}$ is such that $(U_{T+1} - \Pi_S(A_T)) \cdot (A_T - \Pi_S(A_T)) \leq 0$



Then $d(A_T, S) \rightarrow 0$

Intuition: Always walking “towards” the tangent hyperplane with enough “power”

Approachability in Vectors in Expectation

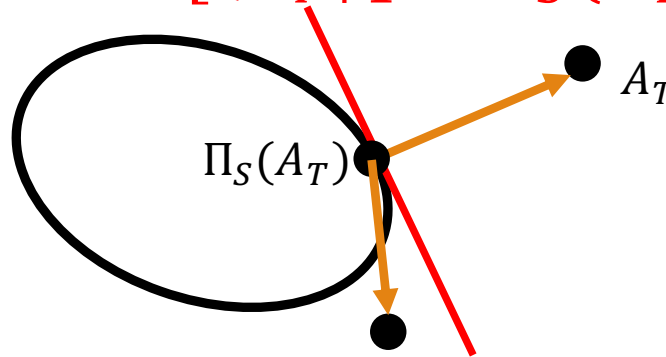
Sequence of **bounded** random vectors $\{U_t\}$, $U_t \in \mathbb{R}^K$

Let average be $A_T = \frac{1}{T} \sum_{t=1}^T U_t$

Let $S \in \mathbb{R}$ be a **convex target set**.

- Let $\Pi_S(A_t)$ be the closest point (projection) of A_t onto S

Assume $\{U_t\}$ is such that $E[(U_{T+1} - \Pi_S(A_T)) \cdot (A_T - \Pi_S(A_T))] \leq 0$



Then $d(A_T, S) \rightarrow 0$ **almost surely** U_{T+1}

U_t 's do not have to be iid. In fact, the expectation doesn't even have to be conditioned on the past!

Blackwell Approachability Game

First, P1 selects action $x_t \in \mathcal{X}$

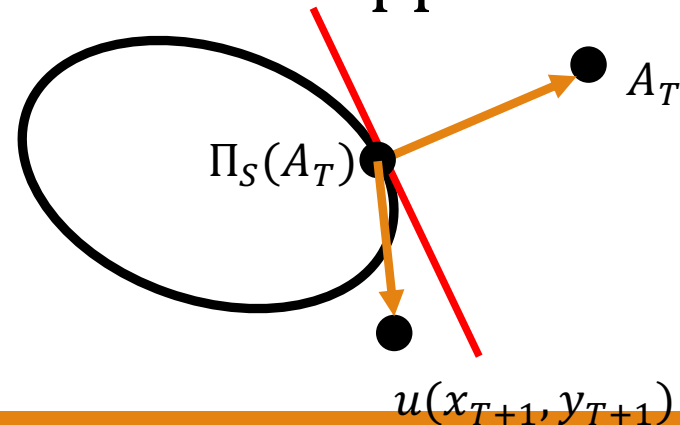
Then, P2 selects action $y_t \in \mathcal{Y}$, adversarial w.r.t. all x_t thus far

P1 incurs a **vector-valued** payoff $u(x_t, y_t)$. Typically, u is biaffine.

P1's goal is to force the average u 's to converge to target set S

$$\min_{\hat{s} \in S} \left\| \hat{s} - \frac{1}{T} \sum_{t=1}^T u(x_t, y_t) \right\| \rightarrow 0 \text{ as } T \rightarrow \infty$$

Idea: Let's use Blackwell approachability



*Want to be able to choose x_T such that no matter how y_{T+1} is chosen, $u(x_{T+1}, y_{T+1})$ will always be on left side of hyperplane!

Forcing Halfspaces and Actions

Convex sets can be difficult to deal with: let's work with halfspaces

Let's consider halfspaces tangent to S : call it \mathcal{H}

$$\mathcal{H} = \{x \in \mathbb{R}^K \mid a^T x \leq b\}$$

\mathcal{H} is forceable if there exists a strategy in x^* such that $u(x^*, y) \in \mathcal{H}$ for all possible choices of y

- x^* is called a **forcing action**

Blackwell: P1's goal will if every halfspace $H \supseteq S$ is forceable

Constructive Proof:

- At T , if $A_T \in S$, choose any $x^* \in \mathcal{X}$
- If not, let \mathcal{H} be halfspace tangent to S containing $\Pi_S(A_T)$, choose x^* to be forcing action of \mathcal{H} .

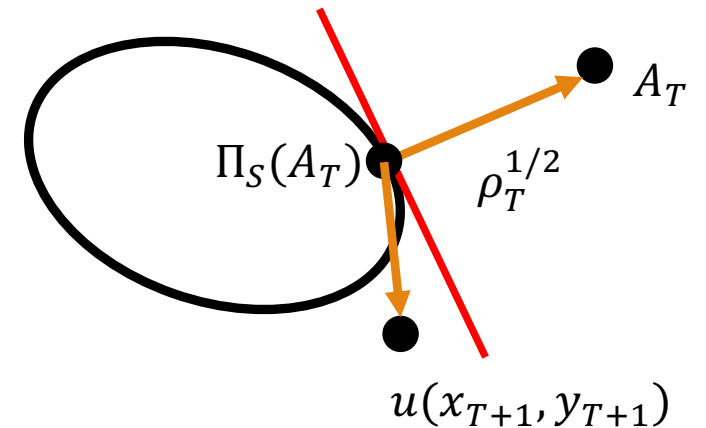
Some derivations (optional)

We could just use Blackwell's theorem, but since this is deterministic it is easy to explicitly show that $d(A_T, S)$ decreases at rate of $1/\sqrt{T}$

$$A_{T+1} = \frac{1}{T+1} \sum_{t=1}^{T+1} u(x_t, y_t) = \frac{T}{T+1} A_T + \frac{1}{T+1} u(x_{T+1}, y_{T+1})$$

$$\rho_T = \|\Pi_S(A_T) - A_T\|^2 = \min_{\hat{s} \in S} \|\hat{s} - A_T\|^2$$

$$\begin{aligned} \rho_{T+1} &= \|\Pi_S(A_{T+1}) - A_{T+1}\|^2 \\ &\leq \|\Pi_S(A_T) - A_{T+1}\|^2 && \text{Projection must be shortest distance} \\ &= \|\Pi_S(A_T) - \frac{T}{T+1} A_T - \frac{1}{T+1} u(x_{T+1}, y_{T+1})\|^2 && \text{Rewrite} \\ &= \left\| \frac{T}{T+1} (\Pi_S(A_T) - A_T) + \frac{1}{T+1} (\Pi_S(A_T) - u(x_{T+1}, y_{T+1})) \right\|^2 && \text{Expand} \\ &= \underbrace{\left(\frac{T}{T+1} \right)^2 \rho_T + \left(\frac{1}{T+1} \right)^2 \|\Pi_S(A_T) - u(x_{T+1}, y_{T+1})\|^2}_{\text{Bounded by Diameter } \Omega^2} + \underbrace{\frac{2T}{(T+1)^2} \langle \Pi_S(A_T) - A_T, \Pi_S(A_T) - u(x_{T+1}, y_{T+1}) \rangle}_{\leq 0 \text{ because forcing action}} \end{aligned}$$

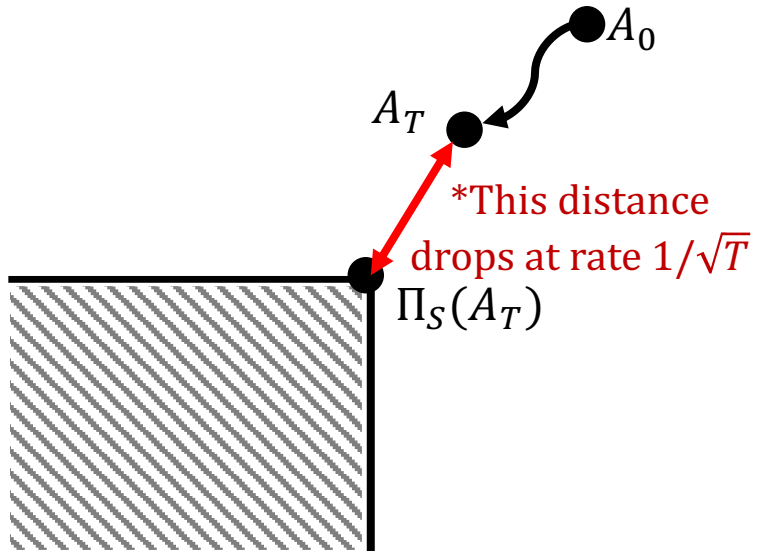


$$(T+1)^2 \rho_{T+1} - T^2 \rho_T \leq \Omega^2 \implies \rho_{T+1} \leq \frac{\Omega^2}{T+1} \implies \min_{\hat{s} \in S} \|\hat{s} - A_T\|_2 \leq \frac{\Omega}{\sqrt{T}}$$

No-regret as a Blackwell Game

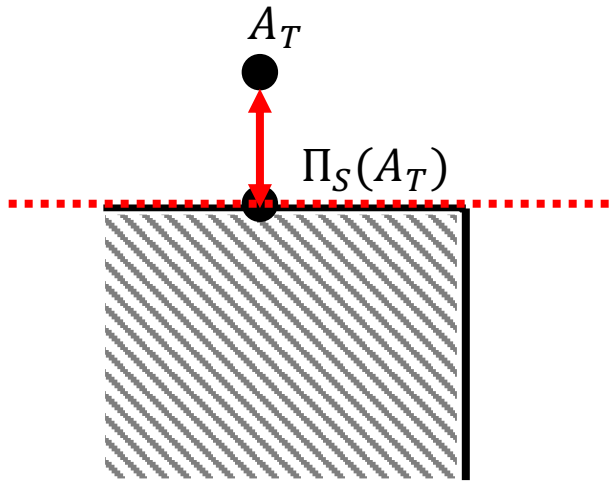
Instantiate

- $u(x_t, y_t) = \ell_t - \langle \ell_t, x_t \rangle$, i.e., regret incurred at t
- Hence, $A_T = \frac{1}{T} \sum_{t=1}^T u(x_t, y_t) = R_T/T$ gives average regret up till T
- $S = \{s \in \mathbb{R}^k | s \leq 0\}$, i.e., nonpositive quadrant
- Hence, if A_T tends to S then we are no-regret (roughly speaking)!

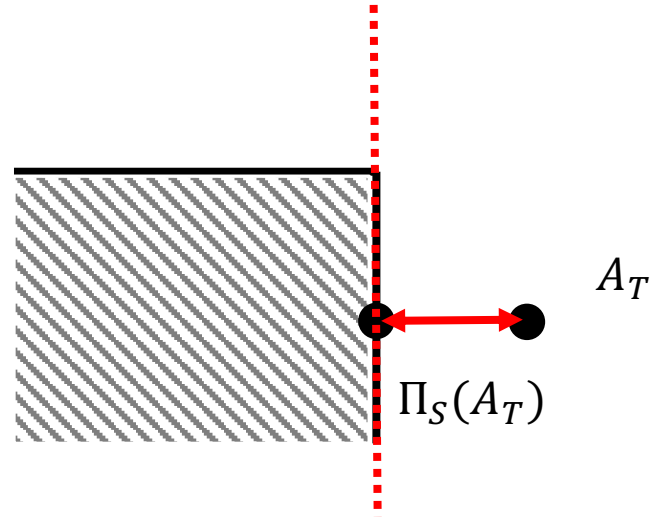


Theorem: The average regret is no greater than $d(A_T, S)$

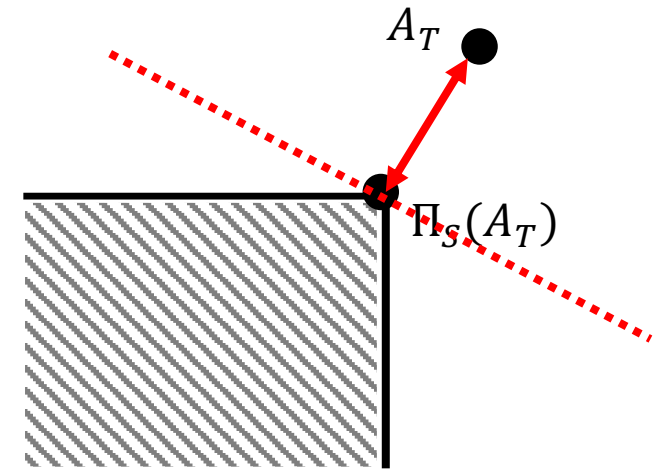
Regret Matching (RM)



Always play action
corresponding to
vertical-axis



Always play action
corresponding to
horizontal-axis



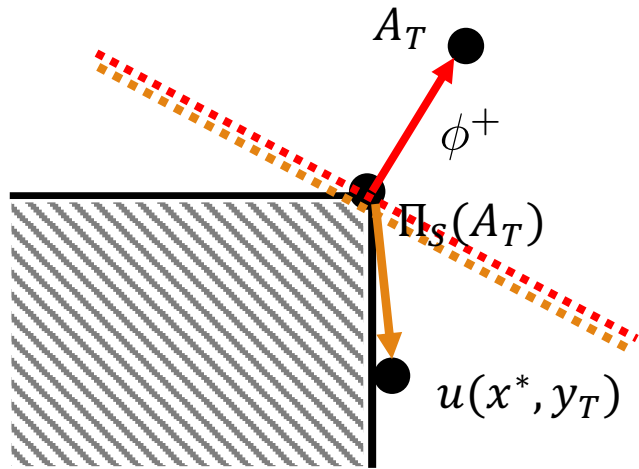
Play according to ratio
of nonnegative
average regrets (?)

Regret Matching Proof

Projection onto nonpositive orthant

$$A_T = [-2, 5, 2, -4] \implies \Pi_S(A_T) = [-2, 0, 0, -4], \underbrace{A_T - \Pi_S(A_T)}_{\triangleq \phi^+} = [0, 5, 2, 0]$$

$$\triangleq \phi^+ \quad \text{*Assume } \neq 0$$



$$\mathcal{H} = \{x \in \mathbb{R}^K \mid \langle \phi^+, x \rangle \leq 0\}$$

$$u(x^*, y_T) \in \mathcal{H} \quad \forall y_T$$

$$\iff \langle \phi^+, u(x^*, y_T) \rangle \leq 0 \quad \forall y_T \quad \text{Definition}$$

$$\iff \langle \phi^+, \ell_T - \langle \ell_T, x^* \rangle \mathbf{1} \rangle \leq 0 \quad \forall \ell_T \in \mathbb{R}^K \quad \text{Definition}$$

$$\iff \langle \phi^+, \ell_T \rangle - \langle \ell_T, x^* \rangle \|\phi^+\|_1 \leq 0 \quad \forall \ell_T \in \mathbb{R}^K \quad \text{Rearrange}$$

$$\iff \langle \ell_T, \frac{\phi^+}{\|\phi^+\|_1} - x^* \rangle \leq 0 \quad \forall \ell_T \in \mathbb{R}^K$$

Forcing action: Just choose $x^* = \frac{\phi^+}{\|\phi^+\|_1}$

RM and RM+

OBSERVEUTILITY(ℓ_t) = reward vector Py_t

$$A_{T+1} = \underbrace{\frac{T}{T+1} A_T}_{\text{Old average regret}} + \underbrace{\frac{1}{T+1} (\ell_T - \langle \ell_T, x_T \rangle 1)}_{\text{Regret to accumulate for this round}}$$

New average regret

RM+: change
average/cumulative
regrets to 0 if negative

NEXTSTRATEGY()

If $\phi^+ = 0$ just choose x^* uniformly at random

$$\underbrace{A_{T+1} = [-2, 5, 2, -4]}_{\text{Average regret}} \implies \underbrace{\phi^+ = [0, 5, 2, 0]}_{\text{Truncate negative regrets}} \implies \underbrace{x^* = [0, 5/7, 2/7, 0]}_{\text{Renormalize}}$$

Note: To make things simpler we could just work with
cumulative regret all the way

Recall: convergence at
rate $1/\sqrt{T}$