Agent-based Systems

Paolo Turrini

www.dcs.warwick.ac.uk/~pturrini p.turrini@warwick.ac.uk

Learning in Games

(Artificial) Poker Stars

Plan for Today

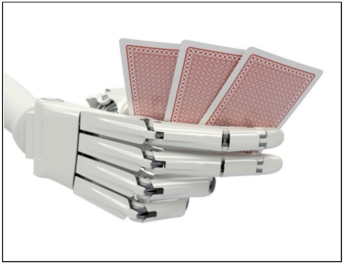
We have seen **extensive games of imperfect information**, where players are typically **uncertainty** about the **current state** of the game being played.

We are going to look at the relevance of this to Artificial Intelligence:

- **Learning through self-play**: the key to many game-playing engines (Alpha Go);
- **Regret**: why it's important to minimise it.

Also we are going to play poker in class.

Poker and Artificial Intelligence



"Robots are unlikely to be welcome in casinos any time soon, especially now that a poker-playing computer has learned to play a virtually perfect game — including bluffing."

(Philip Ball: Game theorists crack poker, *Nature News*, 2015.)

Poker and Artificial Intelligence

- A difficult game
  - Chance, counting odds
  - Bluffing, aggressive play
- Still... a game
  - An extensive game with imperfect information
  - Rational and irrational strategies

What is the right solution concept?

T.W. Sandholm.
Solving Imperfect-Information Games.
Science, 347(6218):122–123, 2015.

An emotional game

- Everyone who has played poker knows how critical "emotions" are in the game;
- Well... it turns out that there is one emotion that computers use better than anyone else;
- This emotion is regret: how bad I've played with respect to how I could have played;
- What is regret really?

Regret in games

- If you play paper and I play paper, my regret for not playing scissors is 1 (the payoff difference!).
- If you play paper and I play rock, my regret for not playing scissors is 2!

	0, 0	−1, 1	1, −1
	1, −1	0, 0	−1, 1
	−1, 1	1, −1	0, 0

Regret in games

- regret for not playing ✊ in (✌, ✋) $= 2$
- regret for not playing ✋ in (✋, ✋) $= 0$
- regret for not playing ✊ in (✋, ✋) $=$ -1

	✊	✋	✌
✊	$0, 0$	$-1, 1$	$1, -1$
✋	$1, -1$	$0, 0$	$-1, 1$
✌	$-1, 1$	$1, -1$	$0, 0$

Let $\langle N, \boldsymbol{A}, \boldsymbol{u} \rangle$ be a normal-form game.

At action profile $\boldsymbol{a}$, the **regret** of player i for not playing a'_i is: $u_i(a'_i, \boldsymbol{a}_{-i}) - u_i(\boldsymbol{a})$.

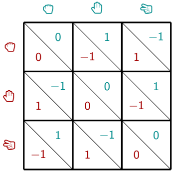
Avoiding feeling bad (without saying it)

How can we use regret to inform future play?

- The idea is that I want to take actions that I wish I had played in the past;
- Obviously if my opponent knew exactly what I was doing it would not be good;
- My strategy needs to be good and **not exploitable**.

How to minimise regret without being predictable?

Regret in games



We do this by **regret matching**: choosing actions at random, with a distribution that is proportional to **positive regrets**.

This means regrets that are proportional to the relative losses one has experienced for not having selected actions in the past.

Regret in games

Games started with (✌, ✋) followed by (✊, ✊):

- regret for not playing ✊ in (✌, ✋) $= 2$
- regret for not playing ✋ in (✊, ✊) $= 1$

	✊	✋	✌
✊	$0, 0$	$-1, 1$	$1, -1$
✋	$1, -1$	$0, 0$	$-1, 1$
✌	$-1, 1$	$1, -1$	$0, 0$

We do this by **regret matching**: choosing actions at random, with a distribution that is proportional to **positive regrets**.

This means regrets that are proportional to the relative losses one has experienced for not having selected actions in the past.

Regret in games

Games started with (✌, ✋) followed by (✊, ✊):

- regret for not playing ✊ in (✌, ✋) $= 2$
- regret for not playing ✋ in (✊, ✊) $= 1$

In the next hand, we choose ✊ with probability $\frac{2}{3}$, ✋ with probability $\frac{1}{3}$, ✌ with probability 0.
Notice: positive regrets divided by their sum.

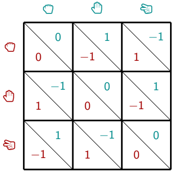
	✊	✋	✌
✊	$0, 0$	$-1, 1$	$1, -1$
✋	$1, -1$	$0, 0$	$-1, 1$
✌	$-1, 1$	$1, -1$	$0, 0$

We do this by **regret matching**: choosing actions at random, with a distribution that is proportional to **positive regrets**.

This means regrets that are proportional to the relative losses one has experienced for not having selected actions in the past.

Cumulating regrets

- Suppose in the next hand I do play $(\frac{2}{3}; \frac{1}{3})$
  … and it turns out to be ;
- Suppose my opponent plays .



Cumulating regrets

- Suppose in the next hand I do play $(\frac{2}{3}$✊$; \frac{1}{3}$✋$)$ … and it turns out to be ✊;
- Suppose my opponent plays ✋.

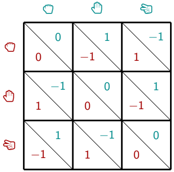
My regret for this hand is 1 for not playing ✋, 2 for not playing ✌, and obviously 0 for not playing ✊.

	✊	✋	✌
✊	0, 0	−1, 1	1, −1
✋	1, −1	0, 0	−1, 1
✌	−1, 1	1, −1	0, 0

Cumulating regrets

- Suppose in the next hand I do play $(\frac{2}{3}; \frac{1}{3})$ … and it turns out to be ;
- Suppose my opponent plays .

My regret for this hand is 1 for not playing , 2 for not playing , and obviously 0 for not playing .



We add the new regrets to the old ones, and play accordingly.

Cumulating regrets

	0, 0	−1, 1	1, −1
	1, −1	0, 0	−1, 1
	−1, 1	1, −1	0, 0

- Suppose in the next hand I do play ($\frac{2}{3}$; $\frac{1}{3}$) ... and it turns out to be ;
- Suppose my opponent plays .

My regret for this hand is 1 for not playing , 2 for not playing , and obviously 0 for not playing .

We have 2 total regrets for , 2 total regrets for , 2 total regrets for .
Our next strategy is going to be ($\frac{2}{6}$; $\frac{2}{6}$; $\frac{2}{6}$)

We add the new regrets to the old ones, and play accordingly.

Cumulating regrets

- Cumulative regrets are good, but not very good.
- If our opponent knows that we are using cumulative regrets, then they are always in a position to best respond.

Can we do better than this? Yes, we can.

- The key idea, as often the case in modern AI, is to play against ourselves;
- We exploit this "hypothetical games" to simulate our opponents and strengthen our strategies.

Learning in Games

Context: you keep playing the same game against the same opponents.

Objective: you want to **learn** their **strategies**.

A good hypothesis might be that the **frequency** with which player i plays action a_i is approximately her **probability** of playing a_i.

Now suppose you always best-respond to those hypothesised strategies. And suppose everyone else does the same. *What will happen?*

We are going to see that for **zero-sum games** this process **converges** to a NE.

This yields a method for **computing a NE** for the (non-repeated) game: just *imagine* players engaging in such "**fictitious play**".

Empirical Mixed Strategies

Given a **history** of actions $H_i^\ell = a_i^0, a_i^1, \dots, a_i^{\ell-1}$ played by player i in ℓ prior plays of game $\langle N, \boldsymbol{A}, \boldsymbol{u} \rangle$, fix her **empirical mixed strategy** $s_i^\ell \in S_i$:

$$s_i^\ell(a_i) \quad = \quad \underbrace{\frac{1}{\ell} \cdot \#\{k < \ell \mid a_i^k = a_i\}}_{\text{relative frequency of } a_i \text{ in } H_i^\ell} \quad \text{for all } a_i \in A_i$$

Best Pure Responses

Recall: Strategy $s_i^\star \in S_i$ is a **best response** for player i to the (partial) strategy profile $\boldsymbol{s}_{-i}$ if $u_i(s_i^\star, \boldsymbol{s}_{-i}) \geqslant u_i(s_i', \boldsymbol{s}_{-i})$ for all $s_i' \in S_i$.
Due to the linearity of expected utilities we get:

Proposition

For any given (partial) strategy profile $\boldsymbol{s}_{-i}$, the set of **best responses** *for player i must include at least one* **pure** *strategy.*

So we can restrict attention to **best pure responses** for player i to $\boldsymbol{s}_{-i}$:

$$a_i^\star \in \operatorname*{argmax}_{a_i \in A_i} u_i(a_i, \boldsymbol{s}_{-i})$$

Fictitious Play

Take any action profile $\boldsymbol{a^0} \in A$ for the normal-form game $\langle N, \boldsymbol{A}, \boldsymbol{u} \rangle$.
Fictitious play of $\langle N, \boldsymbol{A}, \boldsymbol{u} \rangle$, starting in $\boldsymbol{a}^0$, is the following process:

- In round $\ell = 0$, each player $i \in N$ plays action $\mathbf{a_i^0}$.
- In any round $\ell > 0$, each player $i \in N$ plays a **best pure response** to her opponents' **empirical mixed strategies**:

$$\mathbf{a_i^\ell} \in \operatorname*{argmax}_{a_i \in A_i} u_i(a_i, \boldsymbol{s}_{-i}^\ell), \text{ where}$$
$$s_{i'}^\ell(a_{i'}) = \frac{1}{\ell} \cdot \#\{k < \ell \mid a_{i'}^k = a_{i'}\} \text{ for all } i' \in N \text{ and } a_{i'} \in A_{i'}$$

Assume some deterministic way of **breaking ties** between maxima.

This yields a sequence $\boldsymbol{a}^0 \twoheadrightarrow \boldsymbol{a}^1 \twoheadrightarrow \boldsymbol{a}^2 \twoheadrightarrow \ldots$ with a corresponding sequence of empirical-mixed-strategy profiles $\boldsymbol{s}^0 \twoheadrightarrow \boldsymbol{s}^1 \twoheadrightarrow \boldsymbol{s}^2 \twoheadrightarrow \ldots$

Question: Does $\lim_{\ell \to \infty} \boldsymbol{s}^\ell$ exist and is it a meaningful strategy profile?

Example: Matching Pennies

Let's see what happens when we start in the upper lefthand corner HH (and break ties between equally good responses in favour of H):

	H	T
H	$1, -1$	$-1, 1$
T	$-1, 1$	$1, -1$

Any strategy can be represented by a single probability (of playing H).

$$
\begin{aligned}
\text{HH}\ (\tfrac{1}{1}, \tfrac{1}{1}) &\twoheadrightarrow \text{HT}\ (\tfrac{2}{2}, \tfrac{1}{2}) \twoheadrightarrow \text{HT}\ (\tfrac{3}{3}, \tfrac{1}{3}) \twoheadrightarrow \text{TT}\ (\tfrac{3}{4}, \tfrac{1}{4}) \twoheadrightarrow \text{TT}\ (\tfrac{3}{5}, \tfrac{1}{5}) \\
&\twoheadrightarrow \text{TT}\ (\tfrac{3}{6}, \tfrac{1}{6}) \twoheadrightarrow \text{TH}\ (\tfrac{3}{7}, \tfrac{2}{7}) \twoheadrightarrow \text{TH}\ (\tfrac{3}{8}, \tfrac{3}{8}) \twoheadrightarrow \text{TH}\ (\tfrac{3}{9}, \tfrac{4}{9}) \\
&\twoheadrightarrow \text{TH}\ (\tfrac{3}{10}, \tfrac{5}{10}) \twoheadrightarrow \text{HH}\ (\tfrac{4}{11}, \tfrac{6}{11}) \twoheadrightarrow \text{HH}\ (\tfrac{5}{12}, \tfrac{7}{12}) \twoheadrightarrow \cdots
\end{aligned}
$$

Exercise: *Can you guess what this will converge to?*

Convergence Profiles are Nash Equilibria

In general, $\lim_{\ell \to \infty} \boldsymbol{s}^{\ell}$ does not exist (no guaranteed convergence). But:

Lemma

If fictitious play **converges***, then it converges to a* **Nash equilibrium***.*

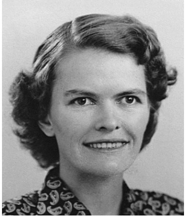
Proof: Suppose $\boldsymbol{s}^{\star} = \lim_{\ell \to \infty} \boldsymbol{s}^{\ell}$ exists. We need to show that $\boldsymbol{s}^{\star}$ is a NE.
To see that it really is, note that $s_i^{\star}$ is the strategy that player i *seems* to be playing, when in fact she best-responds against $\boldsymbol{s}^{\star}{}_{-i}$, which she *believes* to be the profile of strategies of her opponents. ✓
Remark: This lemma is true for arbitrary (not just zero-sum) games.

Convergence for Zero-Sum Games

Good news:

Theorem (Robinson, 1951)

For any **zero-sum game** *and initial action profile,* **fictitious play** *will* **converge** *to a* **Nash equilibrium***.*



Julia Robinson
(1919–1985)

We know that if FP converges, then to a NE.
Thus, we still have to show that it will converge.
The proof of this fact is difficult and we are not going to discuss it here.

J. Robinson.
An Iterative Method of Solving a Game.
Annals of Mathematics, 54(2):296–301, 1951.

Playing against ourselves: the procedure

1. For each player, initialise cumulative regrets to 0;
2. Compute a regret-matching strategy profile;
3. Add the strategy profile played to the strategy profile history;
4. Select each player action profile according the strategy profile;
5. Compute player regrets;
6. Add player regrets to the cumulative regrets;
7. Repeat, for a fixed number of iterations;
8. Return the average strategy profile.

Hart and Mas-Colell (Econometrica, 2000) have shown that this simple procedure converges to a correlated equilibrium, in general, and to the unique NE in two-player zero sum games (like rock-paper-scissors).

Sequential games with imperfect information

- We have all the basics to tackle difficult games;
- The goal is to ‘crack’ poker, remember;
- The idea extend our basic procedure to extensive games with imperfect information.
- I’m only going to present the high-level perspective!

Kuhn Poker

Two players, Ann and Bob, are dealt one of the following cards: $\{A, K, Q\}$.

Ann	Bob	Ann	outcome
pass	pass		$+1$ to higher card
pass	bet	pass	$+1$ to Bob
pass	bet	bet	$+2$ to higher card
bet	pass		$+1$ to Ann
bet	bet		$+2$ to higher card

- Players take turns in starting. So cach player can be playing two different games: the one when they start, and the one when they don't start.
- Each of these games is an extensive game of imperfect information.
- Although Poker is more complicated, it's not that much more complicated really.

Harold E. Kuhn
Simplified two-person poker
Contributions to the Theory of Games, 1950

Let's Play!: Kuhn Poker

Ann	Bob	Ann	outcome
pass	pass		+1 to higher card
pass	bet	pass	+1 to Bob
pass	bet	bet	+ 2 to higher card
bet	pass		+ 1 to Ann
bet	bet		+ 2 to higher card

Let's Play!: Kuhn Poker

Ann	Bob	Ann	outcome
pass	pass		+1 to higher card
pass	bet	pass	+1 to Bob
pass	bet	bet	+ 2 to higher card
bet	pass		+ 1 to Ann
bet	bet		+ 2 to higher card

Scenario 1:
You are Bob. You are dealt Q. Ann passed. What do you do?

Let's Play!: Kuhn Poker

Ann	Bob	Ann	outcome
pass	pass		$+1$ to higher card
pass	bet	pass	$+1$ to Bob
pass	bet	bet	$+ 2$ to higher card
bet	pass		$+ 1$ to Ann
bet	bet		$+ 2$ to higher card

Scenario 1:
You are Bob. You are dealt Q. Ann passed. What do you do?

Scenario 2:
You are Ann. You are dealt Q. What do you do?

Let's Play!: Kuhn Poker

Ann	Bob	Ann	outcome
pass	pass		+1 to higher card
pass	bet	pass	+1 to Bob
pass	bet	bet	+ 2 to higher card
bet	pass		+ 1 to Ann
bet	bet		+ 2 to higher card

Scenario 1:
You are Bob. You are dealt Q. Ann passed. What do you do?

Scenario 2:
You are Ann. You are dealt Q. What do you do?

Scenario 3
You are Ann. You are dealt A. What do you do?

Let's Play!: Kuhn Poker

Ann	Bob	Ann	outcome
pass	pass		$+1$ to higher card
pass	bet	pass	$+1$ to Bob
pass	bet	bet	$+ 2$ to higher card
bet	pass		$+ 1$ to Ann
bet	bet		$+ 2$ to higher card

Scenario 1:
You are Bob. You are dealt Q. Ann passed. What do you do?

Scenario 2:
You are Ann. You are dealt Q. What do you do?

Scenario 3
You are Ann. You are dealt A. What do you do?

Question: What are the objectively bad choices?

CRM

The key tool for AI Poker playing engines is **Counterfactual Regret Minimisation** (usually written with a z instead of an s)

Basically, CRM ...

- represents the game as an extensive game of imperfect information[1]
- uses the regret matching procedure;
- factors in the probabilities of reaching the information sets [2];
- takes into account the fact by moving players learn something about the opponents!

M. Zinkevich et al.
Regret Minimization in Games with Incomplete Information.
NIPS, 2012.

[1] Big games like Texas Hold'em are first 'compressed' into more manageable trees.

[2] In Kuhn Poker, for instance, there are chance nodes (where Nature deals the cards), and decision nodes. Information sets are nodes that a player cannot distinguish. Even for this simple game there are 12 information sets in total.

What we have seen

We have seen how players can "learn" their opponents' strategies.

- Self-play as learning in a repeated game
- Convergence to NE if both players do it. Self-play as learning NE!
- Application to Poker

Next: Look at learning in AI and then get back to GT with the new machinery