# Imperfect-Information Problems

Wen-Yu Lee, Yifu Wang, Qi Yu

# Problem Description

# Problem Description - Why do we need to explore the solution of imperfect-information game

- Perfect Information (pacman, go.....)
  - Nash equilibrium
- Reality: Imperfect Information (financial and energy trading, traffic control)
  - Large scope, expensive
  - Poker

# Approaches

 $\bullet \bullet \bullet$ 

"Deep Reinforcement Learning from Self-Play in Imperfect-Information Games" by J. Heinrich and D. Silver

### Nash Equilibrium

• No players has incentive to change his/her strategy given what the other players are doing.



## Fictitious Self-Play (FSP)

- Fictitious Play (FP): Choose the best response to their opponents' average strategies.
- Fictitious Self-Play (FSP): Mix between their best response and average strategies.



## **Neural Fictitious Self-Play (NFSP)**

• Combination of Neural Network and Fictitious Self-Play

Algorithm 1 Neural Fictitious Self-Play (NFSP) with fitted Q-learning Initialize game  $\Gamma$  and execute an agent via RUNAGENT for each player in the game function RUNAGENT( $\Gamma$ ) Initialize replay memories  $\mathcal{M}_{BL}$  (circular buffer) and  $\mathcal{M}_{SL}$  (reservoir) Initialize average-policy network  $\Pi(s, a \mid \theta^{\Pi})$  with random parameters  $\theta^{\Pi}$ Initialize action-value network  $Q(s, a \mid \theta^Q)$  with random parameters  $\theta^Q$ Initialize target network parameters  $\theta^{Q'} \leftarrow \theta^Q$ Initialize anticipatory parameter  $\eta$ for each episode do Set policy  $\sigma \leftarrow \begin{cases} \epsilon \text{-greedy}(Q), & \text{with probability } \eta \\ \Pi, & \text{with probability } 1 - \eta \end{cases}$ Observe initial information state  $s_1$  and reward  $r_1$ for t = 1, T do Sample action  $a_t$  from policy  $\sigma$ Execute action  $a_t$  in game and observe reward  $r_{t+1}$  and next information state  $s_{t+1}$ Store transition  $(s_t, a_t, r_{t+1}, s_{t+1})$  in reinforcement learning memory  $\mathcal{M}_{BL}$ if agent follows best response policy  $\sigma = \epsilon$ -greedy (Q) then Store behaviour tuple  $(s_t, a_t)$  in supervised learning memory  $\mathcal{M}_{ST}$ end if Update  $\theta^{\Pi}$  with stochastic gradient descent on loss  $\mathcal{L}(\theta^{\Pi}) = \mathbb{E}_{(s,a) \sim \mathcal{M}_{SL}} \left[ -\log \Pi(s, a \mid \theta^{\Pi}) \right]$ Update  $\theta^Q$  with stochastic gradient descent on loss  $\mathcal{L}\left(\theta^{Q}\right) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{M}_{RL}}\left[\left(r + \max_{a'}Q(s',a'\,|\,\theta^{Q'}) - Q(s,a\,|\,\theta^{Q})\right)^{2}\right]$ Periodically update target network parameters  $\theta^{Q'} \leftarrow \theta^Q$ end for end for end function

## Results



#### **NFSP Result**

• The NFSP agents learn by playing against themselves, without any explicitly dened prior knowledge. It uses DQN to replay experiences with Q-Learning.



#### **CFR Result**

• CFR is a RL algorithm that analyses the effect of each move played by the agent and attempts to minimize the loss.

### NFSP VS CRF

		1.1		9	Go	10.	8,	<i>to.</i>	2		Jan Y	36.	19	-à.	2	2	2	Y			20	35	
8																							
7																							
6																							
5																							
4																							
3																							
2																							
6	- rew	and 0.00																					
		aru 0.00																					
-		aru 0.00			- 10-																		
27	39-	10.00	r . ?	¥.	i7	1 <sup>50-</sup>	*	3%.	Ŷ	\$P	No. N	e. 6.	65	ଙ	er.	¢9°.	₹¥ .	Va.	\$V	÷	e)	oj <sup>94-</sup>	
	39-	10- 1	r , ?	4 <sup>37</sup>	i2-	¢₽-	Ŷ	37.	P.	ы́Р	No. N	e. 6.	ଟ	ଙ	er i	¢7.	11	10	\$r	40	°,	9 <sup>98-</sup>	
.9	35-	1 <sup>0</sup>	r	12.	27	°₽~	Ŷ	3	~	ы́Р	No. N	r. 9.	61	G	e.	6 <sup>3</sup> .	1.	10.	\$r	-	°	9 <sup>8</sup>	
° ~ °	39-	1 <sup>0</sup>	r 9	1 <sup>27</sup>	12	<i>₽</i> <sup>57</sup>	*	38	*	5F	No. N	r. G.	65	ଟ	e.	¢2.	11 I	10.	\$r	ŝ	°.	9 <sup>96-</sup>	
9 .9	35-1	19"	r y	47.	'i2"	1 <sup>97</sup>	Ŷ	33.	¥.	₩ <sup>₽</sup>	No. N	° 6'	6 <sup>-</sup>	67	¢.	¢.	Υ.	4a.	¢.	95	d <sub>0</sub> .	9 <sup>8-</sup>	
9	35° 1	1971	r	₩.	12	\$ <sup>5</sup>	*	3	*	5V	8 <sup>97</sup> 8	P. 67	6 <sup></sup>	ଟ	er	¢7.	4¥ 11	4 <u>0</u> .	\$F	\$°		9 <sup>8</sup>	
° ~ ° .9 .8 .7	3 <sup>57</sup> 1	10- 1	r , <del>y</del>	12.	12	\$	*	37	P	\$P	8 <sup>0-</sup> 1	e. e.	51-	ଟ	¢.	¢.	~ · ·	40.	*	80	9 <sup>07</sup>	9**	
9 9 9 8 8 6 9 9 9 9 9 9 9 9 9 9 9 9 9 9	- <del>1</del> 9	10" )	r . Y	1 <sup>27</sup>	'i2''	19 <sup>22</sup>	*	35	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	51	8 <sup>0-</sup> 6	P. 69.	<u>6</u> -	G	e"	¢?-	1 I		*	ď	\$ <sup>37</sup>	9 <sup>n.</sup>	
9 9 8 7 6		10"	r , 97	1 <sup>37</sup>	92° .	~	*	3	7	5r	1 <sup>25-</sup> 1	r. G.	S <sup></sup>	67	¢.	6 <sup>3</sup>	₽¥ I	40-	¢.	÷	°,	9 <sup>44</sup>	
9		10.00	r	47.	Ŷ	1. I. I.	*	Ť	7	\$r	k <sup>o-</sup> k	r 97	6.	ଟ	er .	ę.	₹¥ I	10	S.		9 <sup>337</sup>	9°	
^ ~ 9 .8 .6 .5 .4	35	19.13	r , 97		77	¥	Ŷ	Ť	Ÿ	śν.	8 <sup>97</sup> 8	r 9	5	G	67	¢9°	TV I	4a.	\$¥	ø	ĝ.	9°*	
^ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~	- rew	2	Y	Ψ	12	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	32	3	9	\$V	6 <sup>50</sup> 6	r. 9.	6 <sup>-</sup> -	6°	er .	ę,	4°	19	₩ L	Ø	8 <sup>27</sup>	9 <sup>n.</sup>	
^	- rew	~	* *	Ψ	Ŷ	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	32	3	Ÿ	\$V	87 8	p. G.	9 <sup></sup>	67	e.	<i>Q</i> <sup>*</sup>	£¥.	19	\$ <sup>2</sup>	Ð	<u>6</u> 2.	97°	
^	- rew	~	3	<u><u></u></u>	ήZ	₩ 	Ŷ	3	99 	51/	N2. N	p. 65.	§***	Q <sup>2</sup>	¢.		₹¥.	19 ju		<b>e</b>	<u>6</u>	9°	
99 88 65 5 4 3	-уг - гем	۷۹۳ مربع مربع مربع مربع مربع مربع مربع مربع	3	Ψ	Ŷ	Υ <u>ν</u>	Ŷ	1 1	÷	51	18 <sup>07</sup> 18	P. 67	5	\$P 1	Ŷ		**	42.	\$ <sup>2</sup>	¥	<u>6</u> 2"	9°*	
<ul> <li>9</li> <li>8</li> <li>7</li> <li>6</li> <li>5</li> <li>4</li> <li>3</li> <li>2</li> <li>1</li> </ul>	- гем	40° A 70° A 70° A	3	Ψ.	Ÿ.,		*		9	\$F.	18 <sup>27</sup> 18	P 97	5	e P	\$ 		Y.	4 <sup>1</sup> /2 <sup>-</sup>	\$ <sup>2</sup>	<b>e</b>	<u>6</u> "	07	

# Lesson Learned

 $\bullet \bullet \bullet$ 

#### Lesson Learned

- Nash Equilibrium
- Normal form and Extensive form
- Fictitious Self-Play
- Neural Fictitious Self-Play

# Future Work

 $\bullet \bullet \bullet$ 

### Large-Scale Imperfect-Info Games

- Monte Carlo Neural Fictitious Self-Play (MC-NFSP)
- Asynchronous Neural Fictitious Self-Play (ANFSP)
- Other Models .....



