



Algorithmic Game Theory

Learning in Games

Viliam Lisý

Artificial Intelligence Center
Department of Computer Science, Faculty of Electrical Engineering
Czech Technical University in Prague

(Apr 29, 2019)

Plan



Online learning and prediction

single agent learns to select the best action

Learning in normal form games

the same algorithms used by multiple agents

Learning in extensive form games

generalizing these ideas to sequential games

DeepStack



Algorithmic Game Theory

Learning in extensive form games

Viliam Lisý

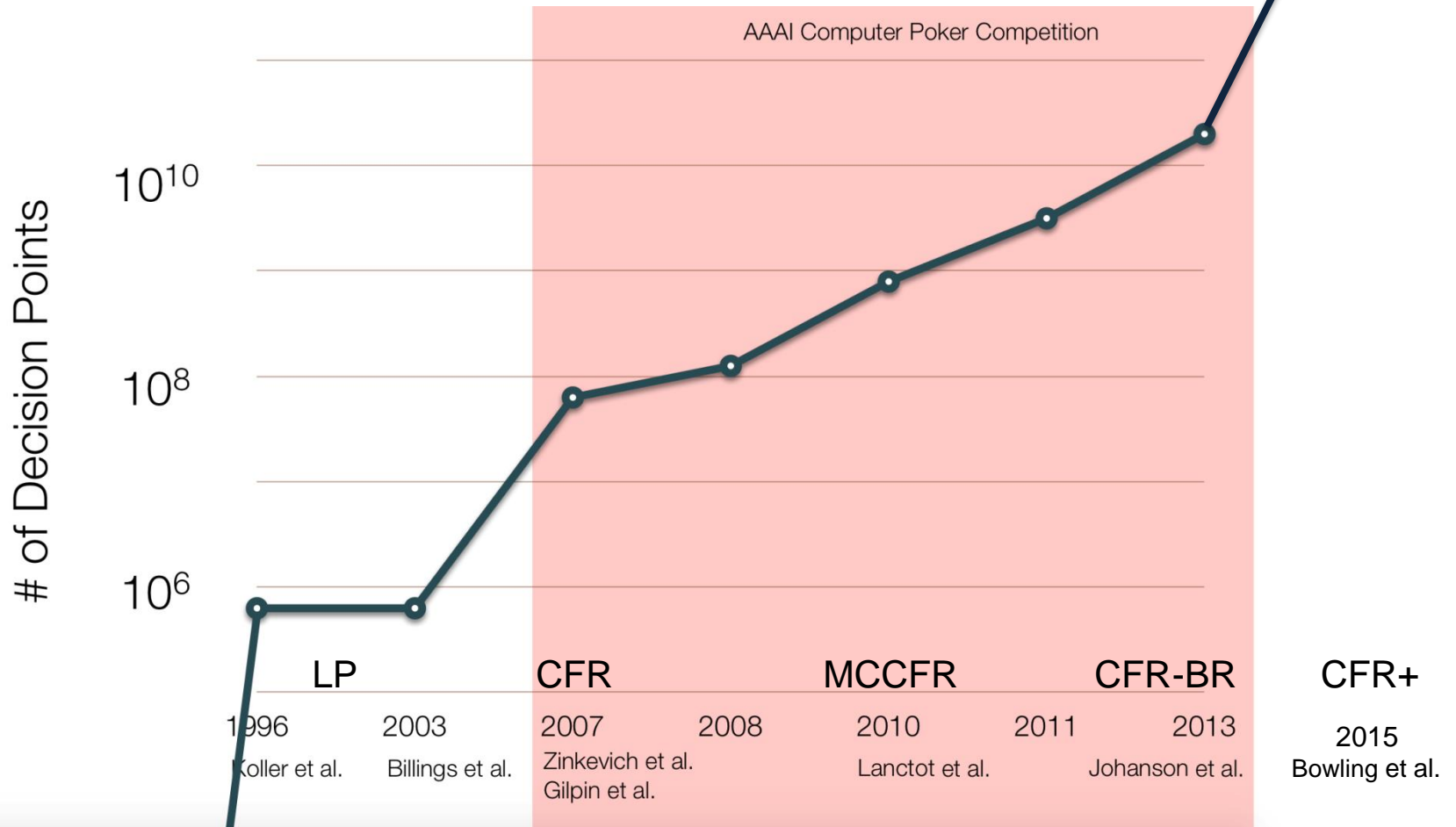
Artificial Intelligence Center
Department of Computer Science, Faculty of Electrical Engineering
Czech Technical University in Prague

(Apr 29, 2019)

Impact on poker performance



1.4×10^{13} Heads-Up Limit Texas Hold'em



Based on M. Bowling's slide from AAI 2015 keynote

Solving games by regret minimization



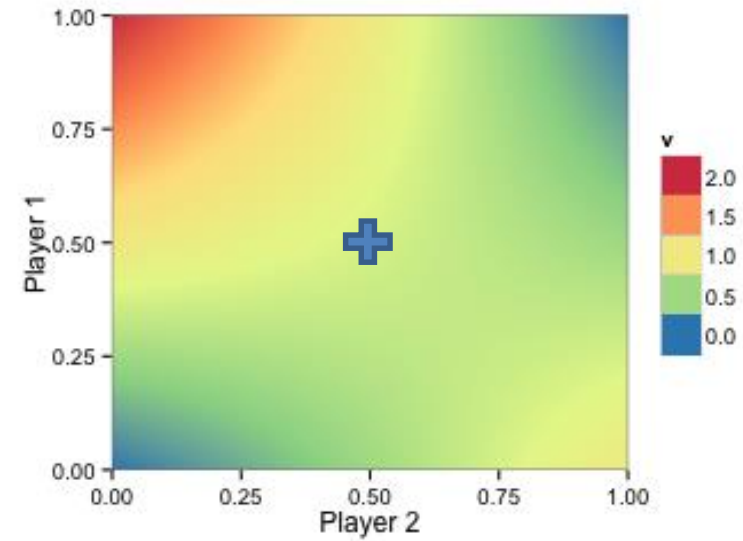
Theorem: If the **average external regret** for each player's sequence of strategies in a zero-sum game is $\bar{r}^T < \epsilon$ then the average strategies $\bar{\sigma}^T = \frac{1}{T} \sum_{t=0}^T \sigma^t$ form an 2ϵ -Nash equilibrium

Regret matching+



σ^t

	0.5	0.5
0.5	2	0
0.5	0	1



Regret matching+



Iteration:

1

$\overline{\sigma}_2$

R_2

r_2

σ^t

0	0

0.5 0.5

2	0
0	1

$\overline{\sigma}_1$

R_1

r_1

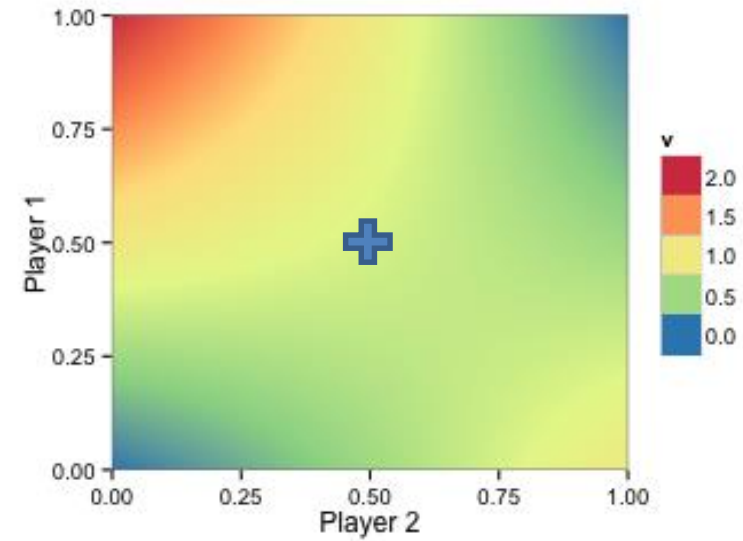
	0
	0

0.25

-0.25

0.5

0.5



Regret matching+



Iteration:

1

$\bar{\sigma}_1$	R_1
	0.25
	0

r_1	σ^t
0.25	0.5
-0.25	0.5

$\bar{\sigma}_2$

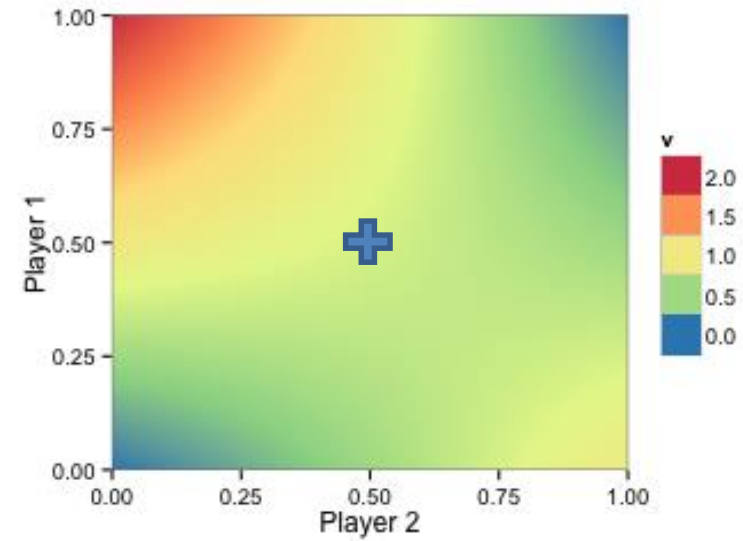
R_2

r_2

σ^t

0	0

0.5	0.5
2	0
0	1



Regret matching+



Iteration:

1

$\bar{\sigma}_1$	R_1
	0.25
	0

0.25

-0.25

$\bar{\sigma}_2$

R_2

r_2

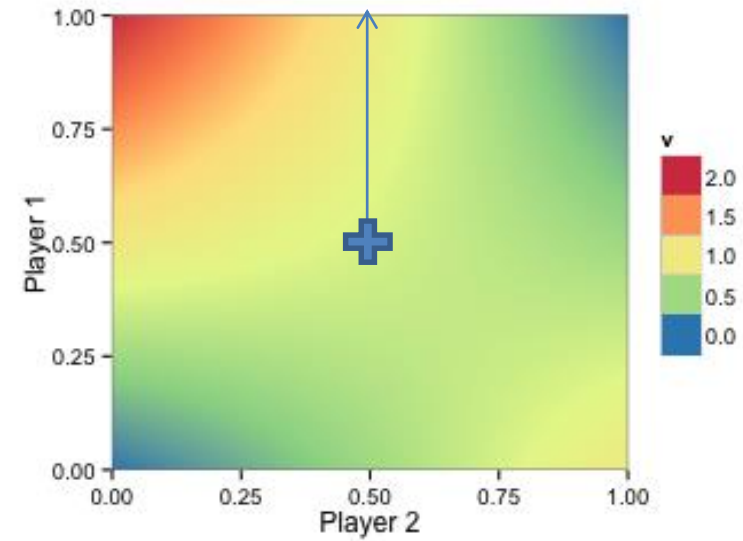
σ^t

0	0

0.5

0.5

2	0
0	1



Regret matching+



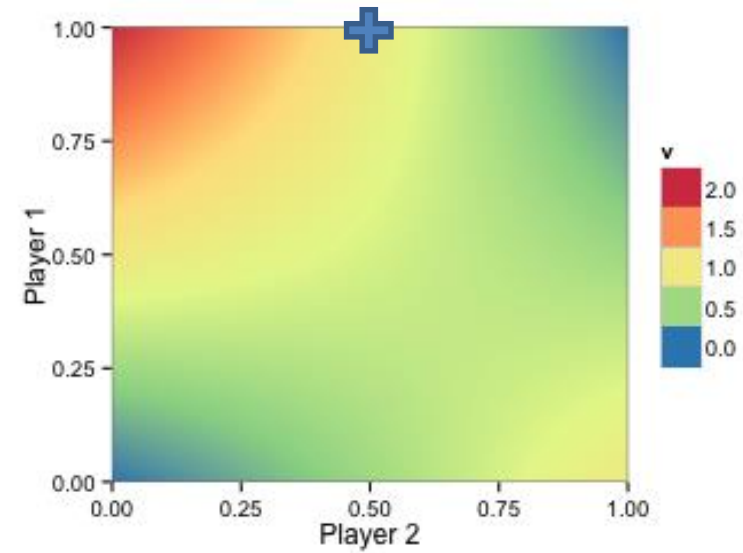
Iteration:

1

$\bar{\sigma}_1$	R_1
1	0.25
0	0

r_1	σ^t
1	1
0	0

$\bar{\sigma}_2$		
R_2	0	0
r_2	-1	1
σ^t	0.5	0.5
1	2	0
0	0	1



Regret matching+



Iteration:

1

$\overline{\sigma}_2$

R_2

r_2

σ^t

0	1

2	0
0	1

$\overline{\sigma}_1$

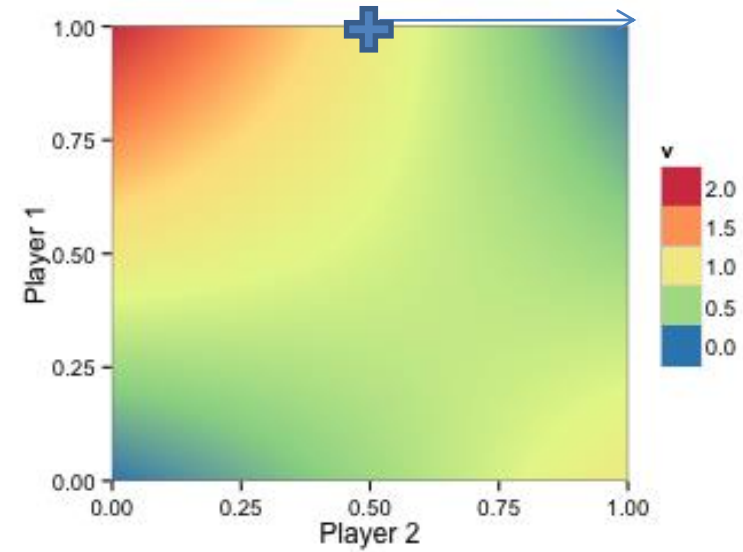
R_1

r_1

1

0

1	0.25
0	0



Regret matching+



Iteration:
2

$\bar{\sigma}_1$	R_1
1	0.25
0	0

r_1	0	1
σ^t	1	0

$\bar{\sigma}_2$

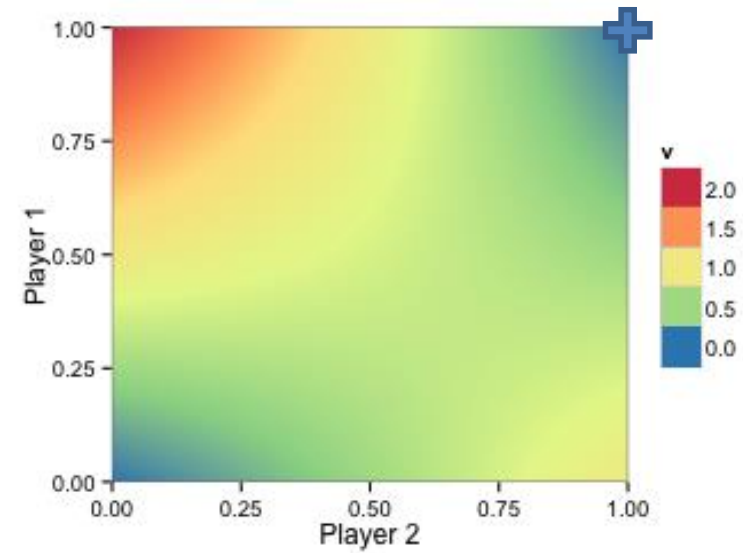
R_2

r_2

σ^t

0	1
0	1

2	0
0	1



Regret matching+



Iteration:
2

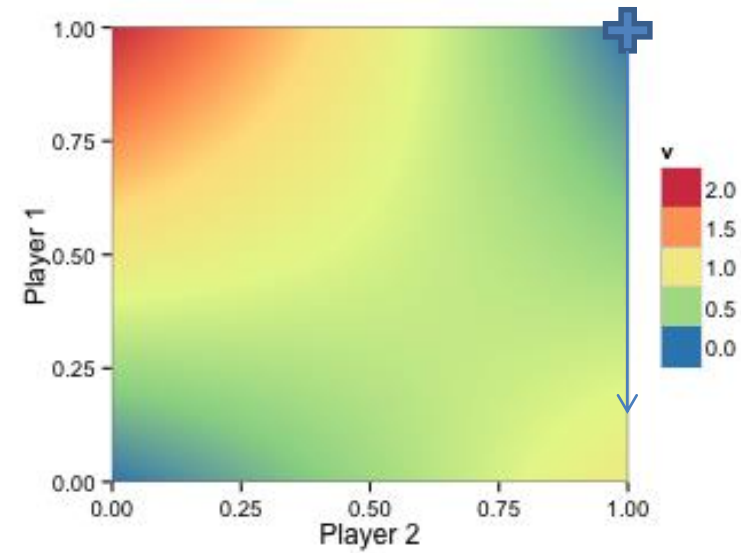
$\bar{\sigma}_1$	R_1
1	0.25
0	1

r_1
0
1

$\bar{\sigma}_2$
 R_2
 r_2
 σ^t

0	1
0	1

2	0
0	1



Regret matching+



Iteration:
2

$\bar{\sigma}_1$	R_1
0.46	0.25
0.54	1

r_1
0
1

σ^t
0.2
0.8

$\bar{\sigma}_2$

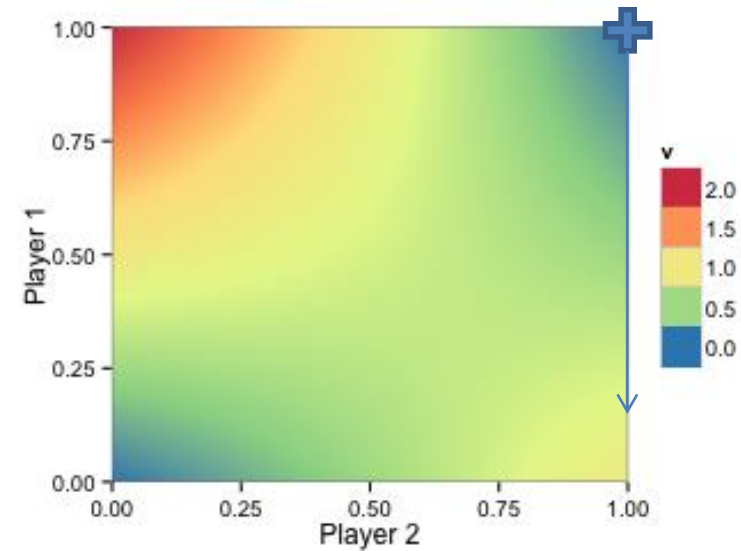
R_2

r_2

σ^t

0	1
0	1

2	0
0	1



Regret matching+



Iteration:
2

$\bar{\sigma}_1$	R_1
0.46	0.25
0.54	1

r_1

$\bar{\sigma}_2$

R_2

r_2

σ^t

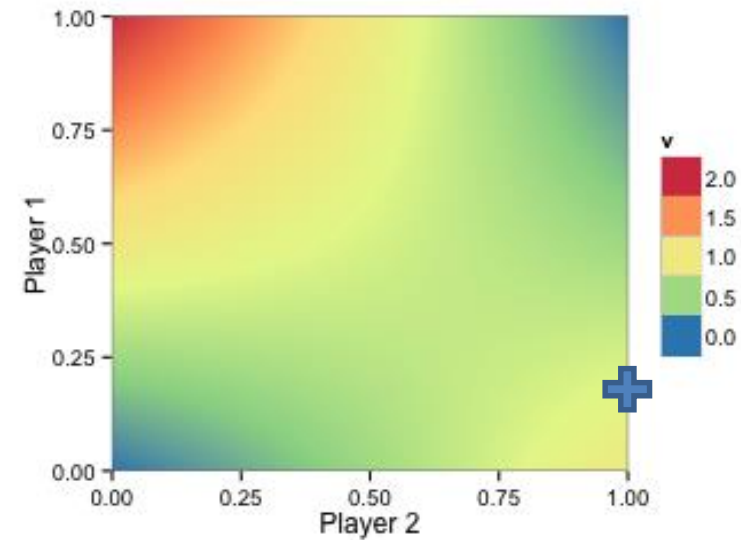
0.2

0.8

0	1
0	1

0.4 0

2	0
0	1



Regret matching+



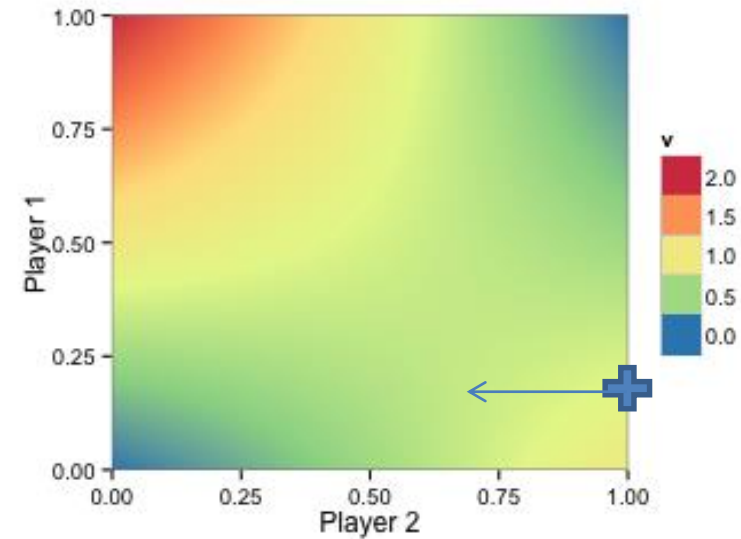
Iteration:
2

$\bar{\sigma}_1$	R_1
0.46	0.25
0.54	1

r_1	σ^t
0.2	2
0.8	0

$\bar{\sigma}_2$	0	1
R_2	0.4	1
r_2	0.4	0

2	0
0	1



Regret matching+



Iteration:
2

$\bar{\sigma}_1$	R_1
0.46	0.25
0.54	1

r_1

$\bar{\sigma}_2$

R_2

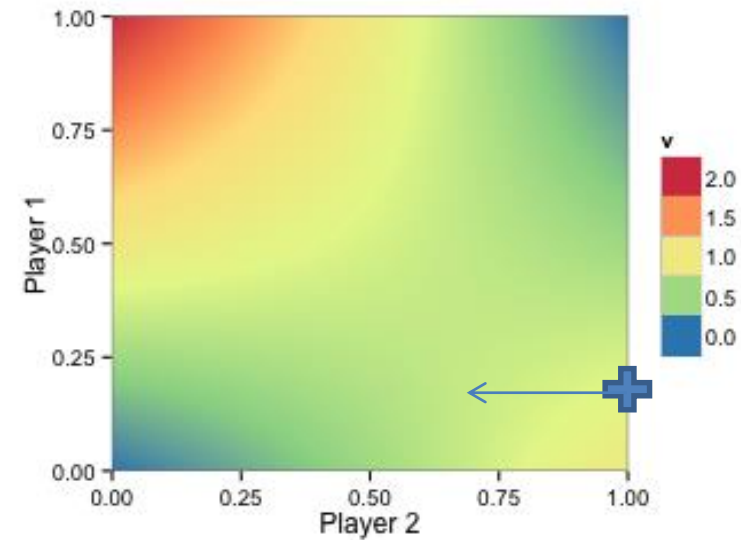
r_2

σ^t

0.2

0.8

0.19	0.81
0.4	1
0.4	0
0.29	0.71
2	0
0	1

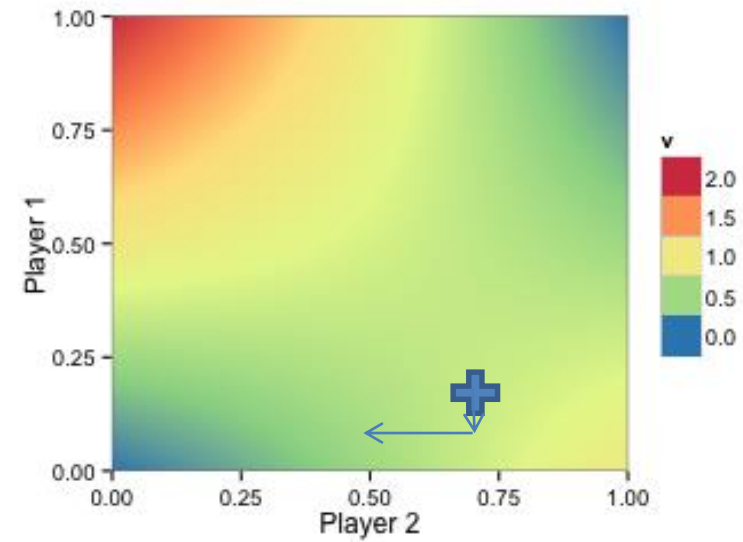


Regret matching+



Iteration:

3

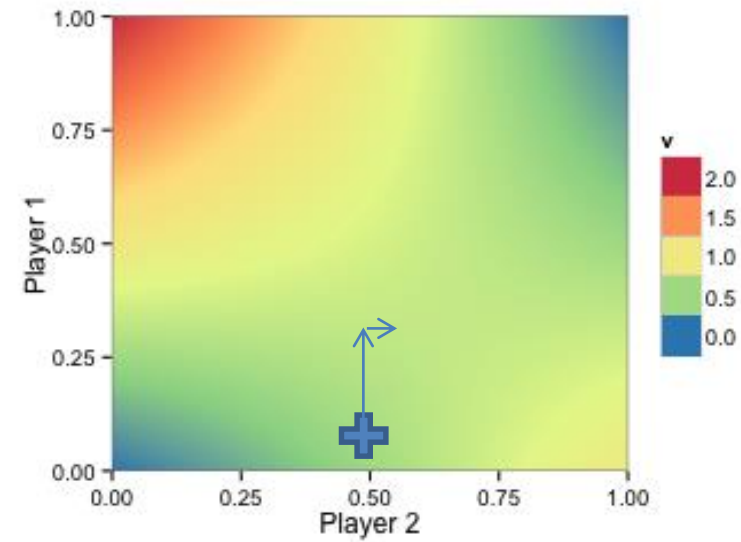


Regret matching+



Iteration:

4

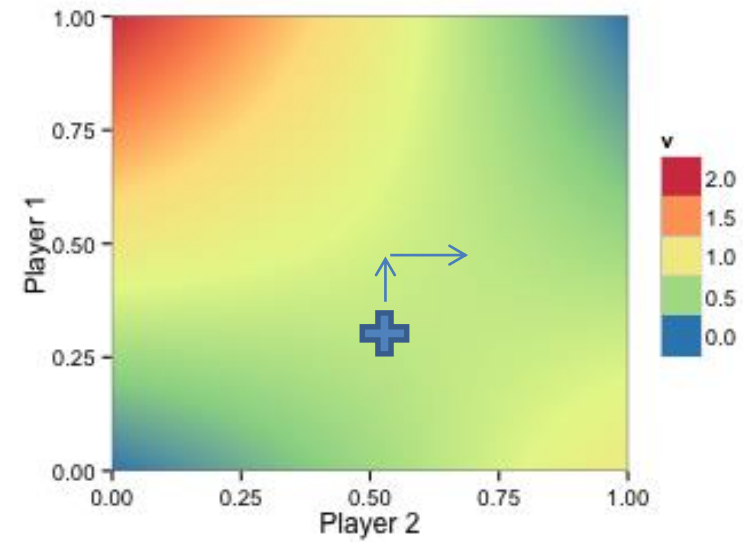


Regret matching+



Iteration:

5

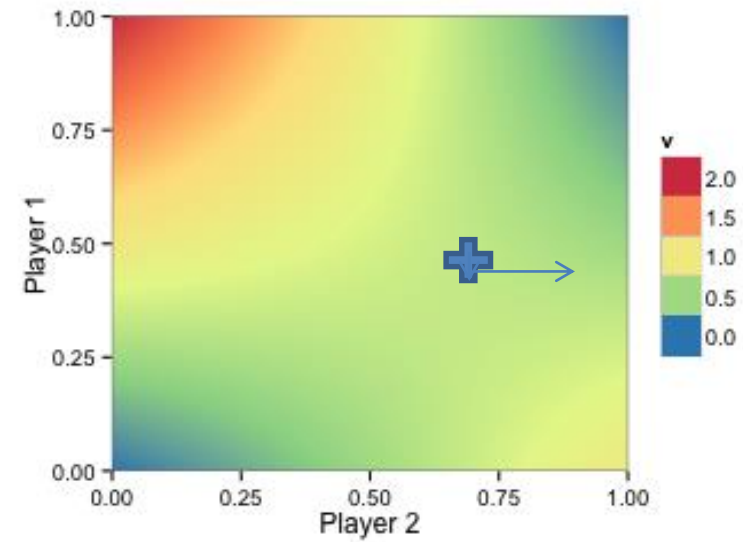


Regret matching+



Iteration:

6

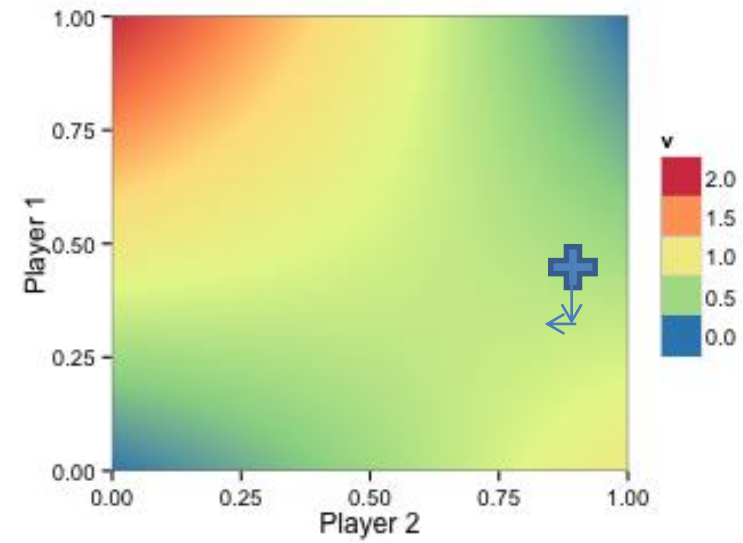


Regret matching+



Iteration:

7

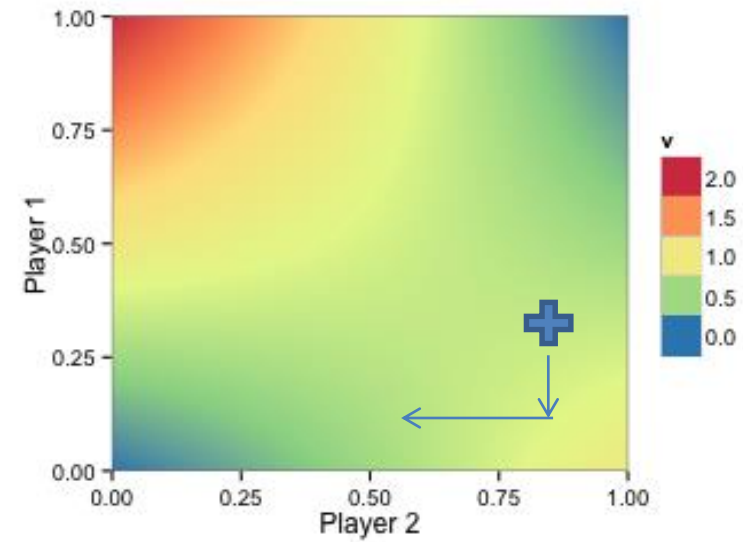


Regret matching+



Iteration:

8



Regret matching+



Iteration:
8

$\bar{\sigma}_1$	R_1
0.33	0.17
0.67	1.30

r_1

$\bar{\sigma}_2$

R_2

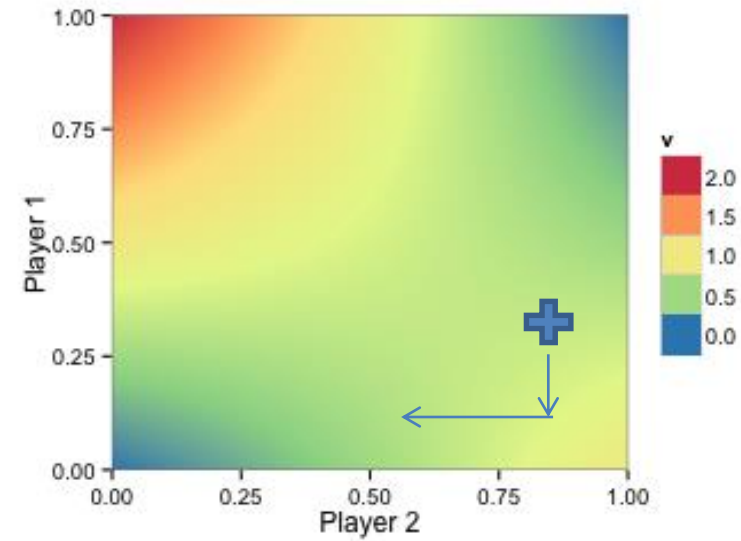
r_2

σ^t

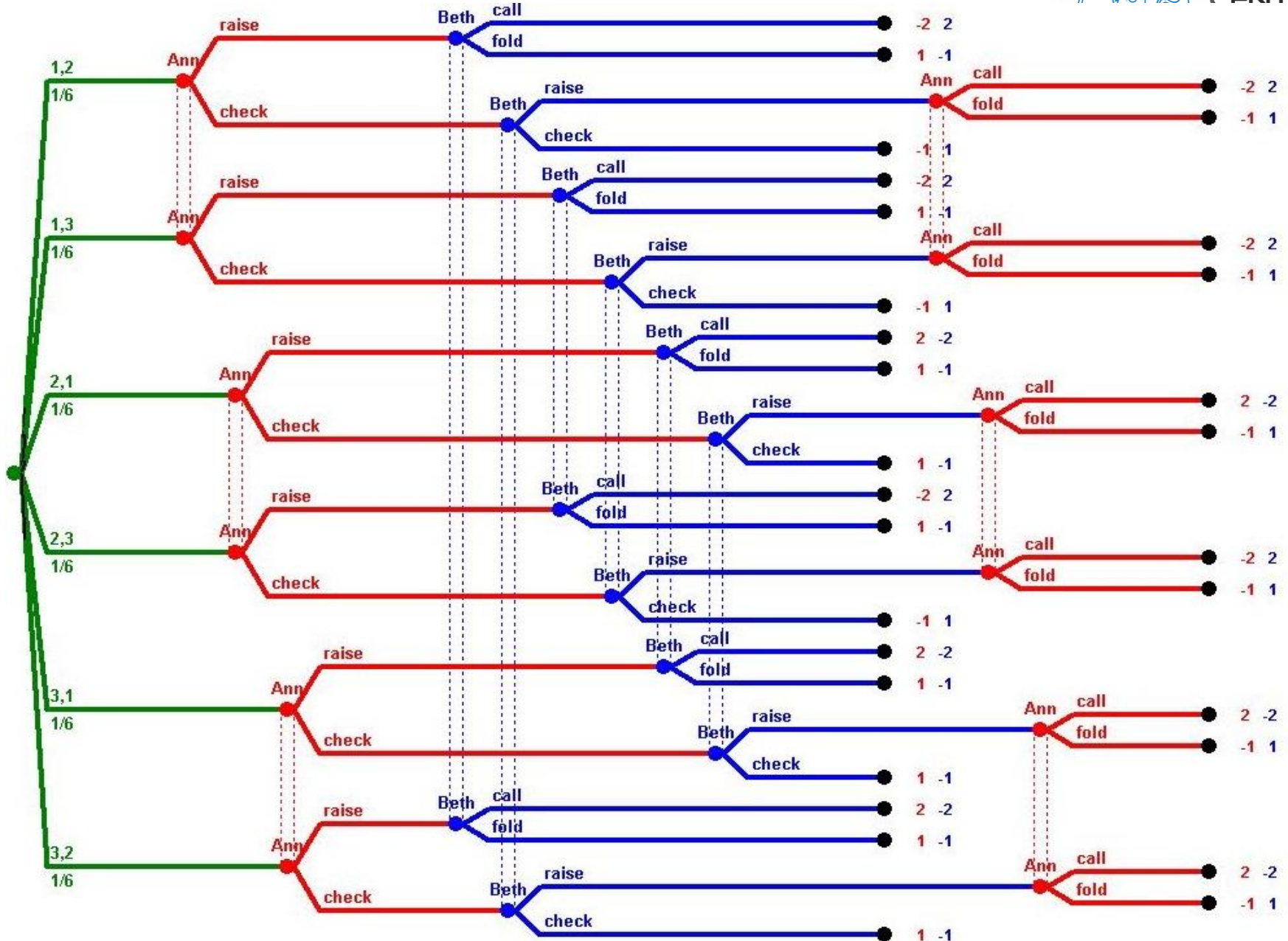
0.11

0.88

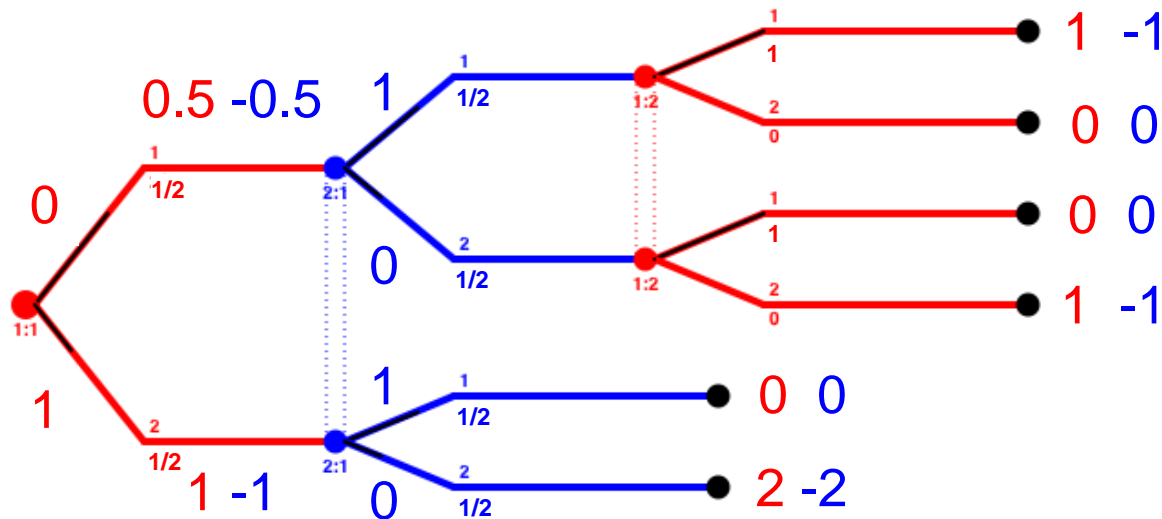
0.30	0.70
0.83	1.15
0.42	0.58
2	0
0	1



Extensive form games



Counterfactual Regret - Motivation



1	0
0	1

0	2
---	---

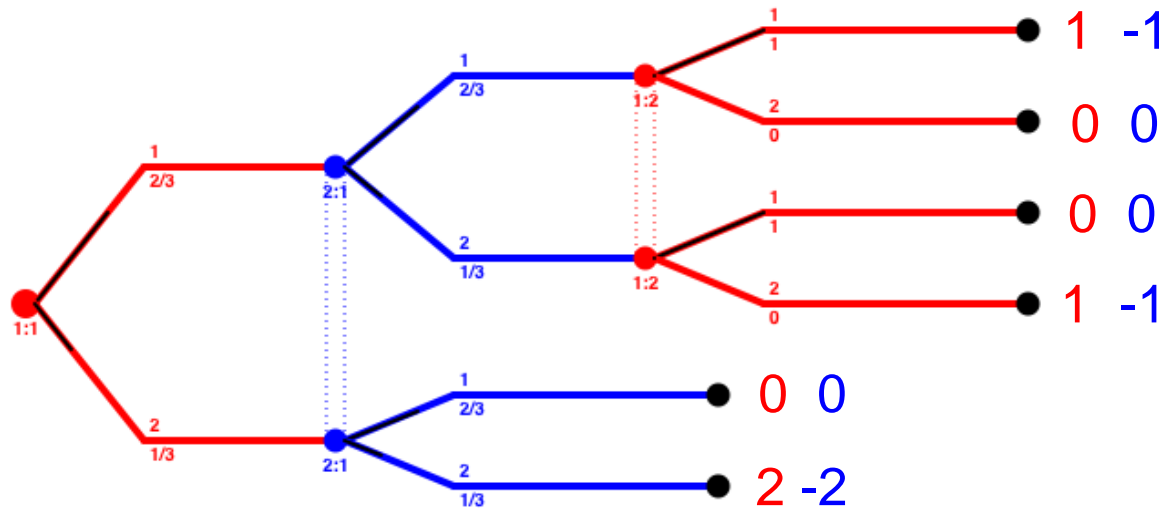
Take the current reach probabilities?

-> undefined belief

Take only opponent's reach probability!

-> defined where necessary

Counterfactual Regret - Definition



Counterfactual value: $v_i^\sigma(I, a) = \sum_{(h,z) \in Z_I} \pi_{-i}^\sigma(h) \pi^\sigma(ha, z) u_i(z)$

Counterfactual regret: $r^t(I, a) = v_i^{\sigma^t}(I, a) - v_i^{\sigma^t}(I)$

Can be computed in **one tree walk**

Counterfactual Regret Minimization

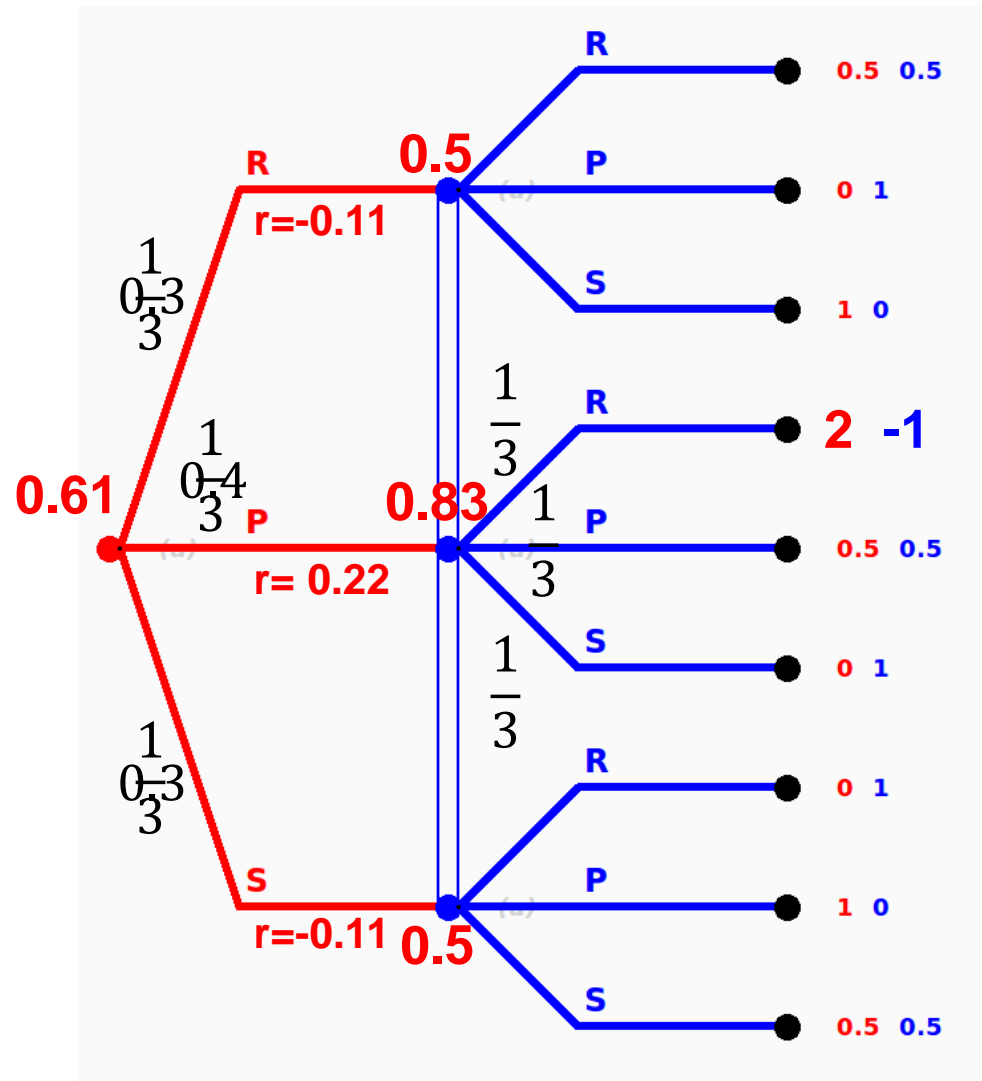


- 1) Walk the tree to compute counterfactual values in all ISs
- 2) Use RM, RM+, Hedge,... to compute next strategy for each IS
- 3) Goto 1

- 4) Return **mean** of all used strategies

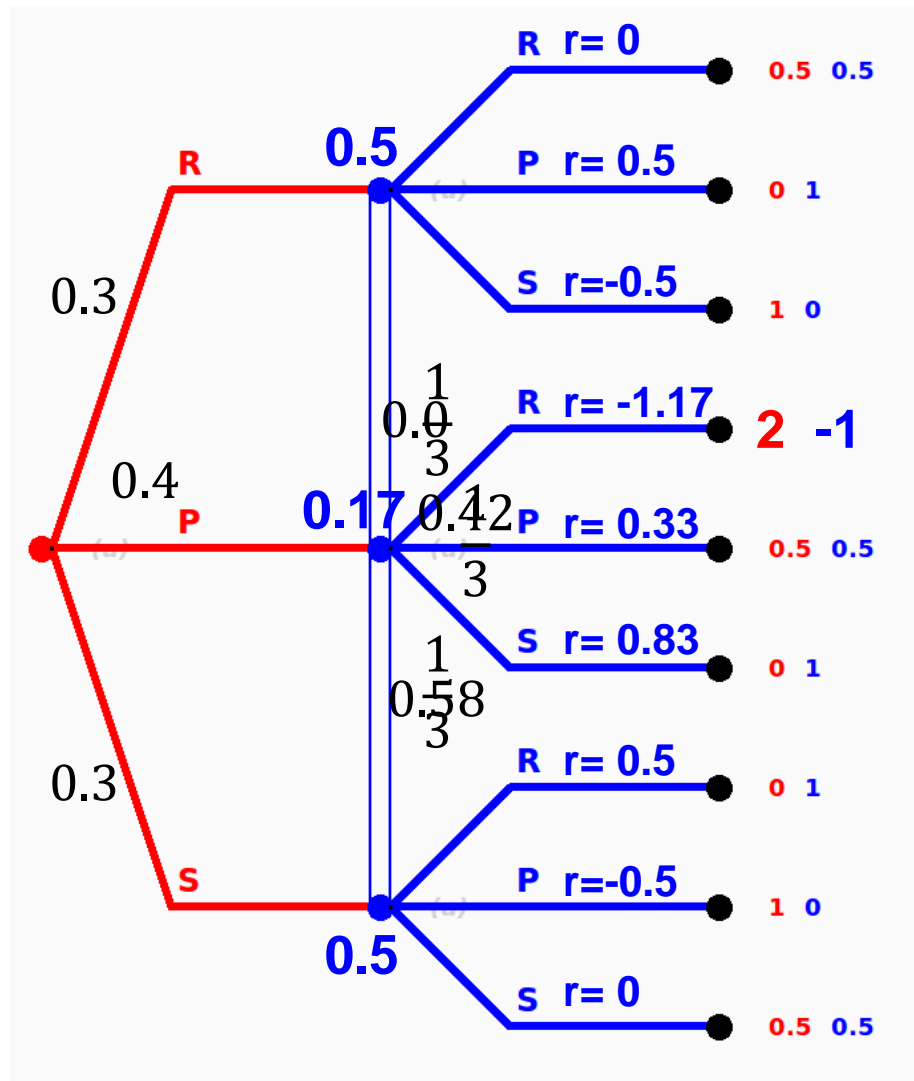
Counterfactual regret minimization

Player 1 iteration



Counterfactual regret minimization

Player 2 iteration



$$\begin{aligned}
 R: & 0.3 \cdot 0 \\
 & + 0.4 \cdot -1.17 \\
 & + 0.3 \cdot 0.5 \\
 & = -0.318
 \end{aligned}$$

$$\begin{aligned}
 P: & 0.3 \cdot 0.5 \\
 & + 0.4 \cdot 0.33 \\
 & + 0.3 \cdot -0.5 \\
 & = 0.132
 \end{aligned}$$

$$\begin{aligned}
 S: & 0.3 \cdot -0.5 \\
 & + 0.4 \cdot 0.83 \\
 & + 0.3 \cdot 0 \\
 & = 0.182
 \end{aligned}$$

Counterfactual Regret Minimization



Theorem (Zinkevich et al. 2008): For a sequence of (mixed) strategies σ_i^t , let $R_{i,imm}^T(I) = \max_a \sum_{t \in 1..T} r^t(I, a)$ then

$$R_{i,full}^T \leq \sum_I R_{i,imm}^{T,+}(I)$$

Proof: Let $D(I)$ be the information sets reachable from I , $Succ_i(I, a)$ be the possible next information sets, $Succ_i(I) = \cup_{a \in A(I)} Succ_i(I, a)$.

$$R_{i,full}^T(I) = \max_{\sigma' \in \Sigma_i} \sum_{t \in 1..T} \left(v_i(\sigma^t |_{D(I) \rightarrow \sigma'}, I) - v_i(\sigma^t, I) \right)$$

$$v_i^\sigma(I, a) = \sum_{(h,z) \in Z_I} \pi_{-i}^\sigma(h) \pi^\sigma(ha, z) u_i(z); \quad r^t(I, a) = v_i^{\sigma^t}(I, a) - v_i^{\sigma^t}(I)$$

$$R_{i,imm}^T(I) = \max_{a \in A(I)} \sum_{t \in 1..T} (v_i(\sigma^t |_{I \rightarrow a}, I) - v_i(\sigma^t, I))$$

Lemma: $R_{i,full}^T(I) \leq R_{i,imm}^T(I) + \sum_{I' \in Succ_i(I)} R_{i,full}^{T,+}(I')$

$$\begin{aligned}
 R_{i,full}^T(I) &= \max_{a \in A(I)} \max_{\sigma' \in \Sigma_i} \sum_{t \in 1..T} \\
 &\quad (v_i(\sigma^t|_{I \rightarrow a}, I) - v_i(\sigma^t, I)) \\
 &\quad + \sum_{I' \in Succ_i(I,a)} succ_i^\sigma(I'|I, a) \left(\frac{\pi_{-i}^{\sigma^t}(I)}{\pi_{-i}^{\sigma^t}(I')} \right) (v_i(\sigma^t|_{D(I) \rightarrow \sigma'}, I') - v_i(\sigma^t, I')) \\
 R_{i,full}^T(I) &\leq \max_{a \in A(I)} \max_{\sigma' \in \Sigma_i} \sum_{t \in 1..T} (v_i(\sigma^t|_{I \rightarrow a}, I) - v_i(\sigma^t, I)) \\
 &\quad + \max_{a \in A(I)} \max_{\sigma' \in \Sigma_i} \sum_{t \in 1..T} \sum_{I' \in Succ_i(I,a)} (v_i(\sigma^t|_{D(I') \rightarrow \sigma'}, I') - v_i(\sigma^t, I'))
 \end{aligned}$$

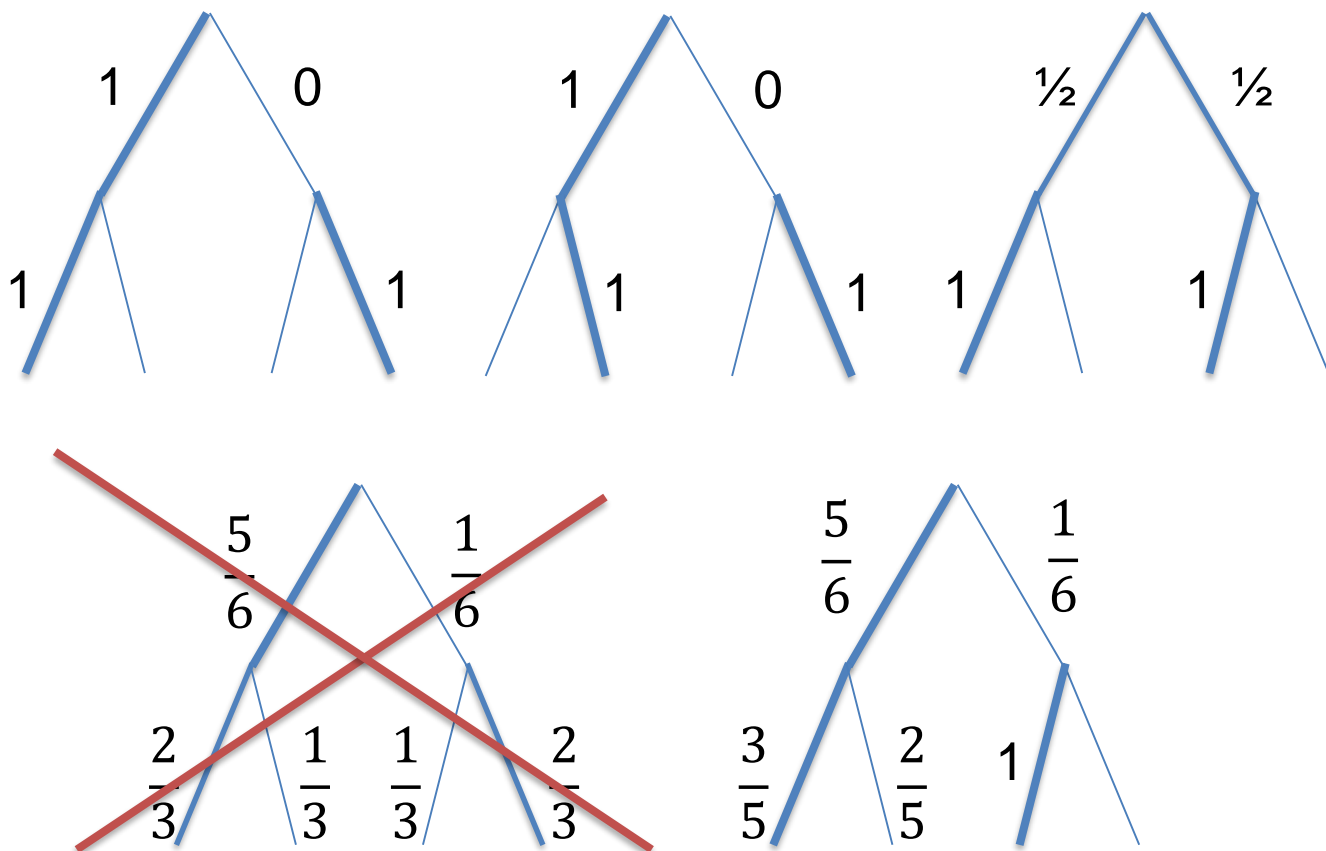
$$\begin{aligned}
 R_{i,full}^T(I) &\leq R_{i,imm}^T(I) + \max_{a \in A(I)} \sum_{I' \in Succ_i(I,a)} R_{i,full}^T(I') \\
 &\leq R_{i,imm}^T(I) + \sum_{I' \in Succ_i(I)} R_{i,full}^{T,+}(I').
 \end{aligned}$$

The proof of the theorem is completed by induction, using the Lemma above.

Average Strategy in CFR



$$\bar{\sigma}_i^T(I, a) = \frac{\sum_{t=1}^T \pi_i^{\sigma^t}(I) \sigma^t(I, a)}{\sum_{t=1}^T \pi_i^{\sigma^t}(I)}$$



CFR+ Convergence Speed

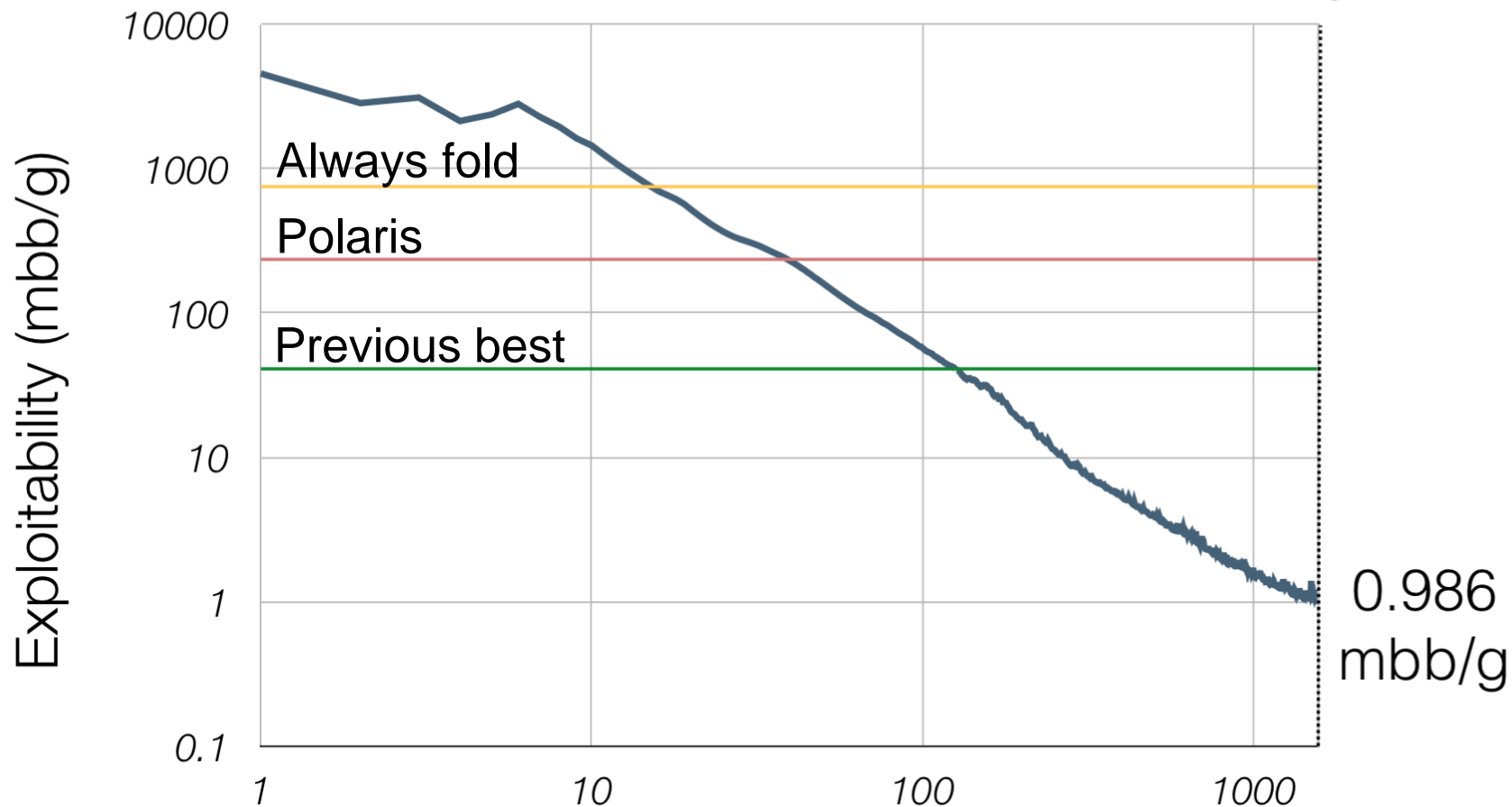


Theorem (Tammelin et al. 2015): The mean strategies from CFR+ in a game with payoff range Δ , $A = \max_I |A(I)|$, after T iterations form an $\frac{2(|I_1|+|I_2|)\Delta\sqrt{A}}{\sqrt{T}}$ -Nash equilibrium.

Solving Limit Texas Hold'em (Bowling et al., Science 2015)



69 days
900 CPU-years



CFR Variants – MCCFR



Monte Carlo Counterfactual Regret Minimization (Lanctot et al. 2009)

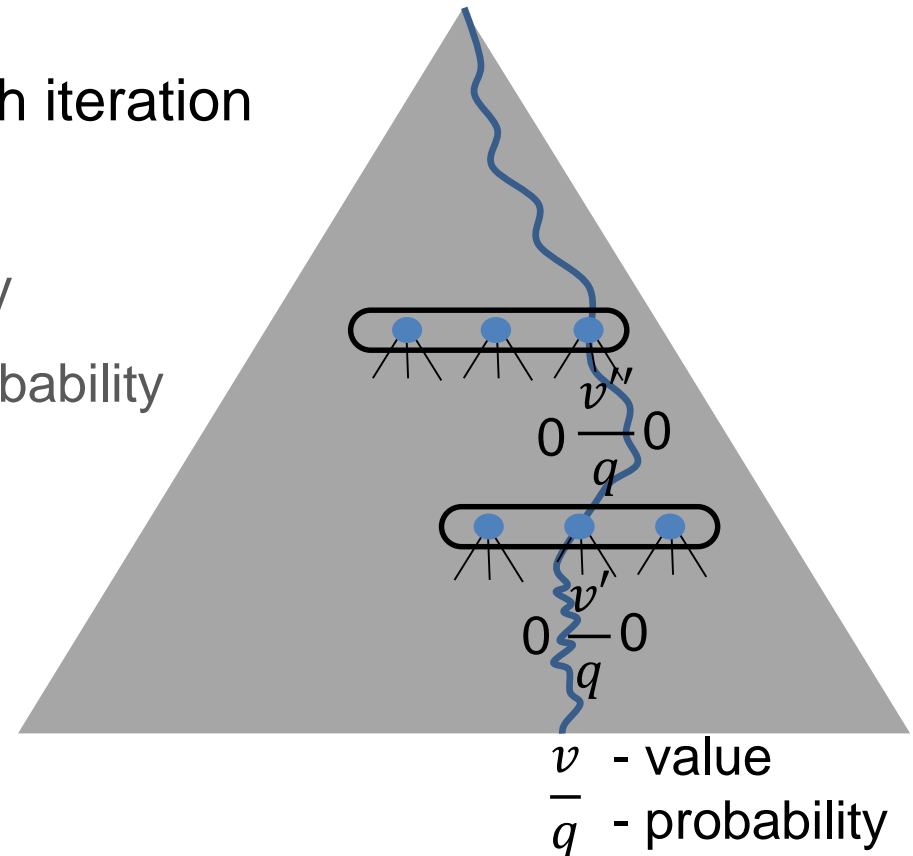
Samples a subset of tree in each iteration

Importance sampling trick

Unbiased estimator of real CFV

Still need to weight by opp. probability

Domain specific sampling



MCCFR



Recall CFV: $v_i^\sigma(I) = \sum_{(h,z) \in Z_I} \pi_{-i}^\sigma(h) \pi^\sigma(h, z) u_i(z)$

Let $Q = \{Q_1, Q_2, \dots, Q_{|Q|}\}$ blocks in Z such that $\bigcup_{Q_j \in Q} Q_j = Z$

MCCFR samples $Q_j \in Q$ with probability q_j . Let $q(z) = \sum_{j: z \in Q_j} q_j$.

$$\tilde{v}_i^\sigma(I|j) = \sum_{(h,z) \in Q_j \cap Z_I} \frac{1}{q(z)} \pi_{-i}^\sigma(h) \pi^\sigma(h, z) u_i(z)$$

Sampling schemes:

outcome sampling

external sampling

Lemma: $\mathbf{E}_j[\tilde{v}_i^\sigma(I|j)] = v_i^\sigma(I)$

Proof: $\mathbf{E}_j[\tilde{v}_i^\sigma(I|j)] = \sum_j q_j \sum_{(h,z) \in Q_j \cap Z_I} \frac{1}{q(z)} \pi_{-i}^\sigma(h) \pi^\sigma(h, z) u_i(z) = v_i^\sigma(I)$

MCCFR Convergence Bound



Theorem (simplified): For any $p \in (0,1]$ and $Q_j \in \mathcal{Q}$,

$$\sum_I \left(\sum_{(h,z) \in Q_j \cap Z_I} \frac{\pi^\sigma(h,z) \pi_{-i}^\sigma(h)}{q(z)} \right)^2 \leq \frac{1}{\delta^2}$$

then with probability at least $1 - p$, average overall regret

$$\bar{r}_{i,full}^T \leq \left(1 + \frac{\sqrt{2}}{\sqrt{p}} \right) \frac{1}{\delta} \Delta |\mathcal{I}_i| \frac{\sqrt{|A|}}{\sqrt{T}}$$

MCCFR Convergence Bound



Proof sketch:

Markov's inequality: $P(x \geq a) \leq \frac{\mathbf{E}[x]}{a}$ (for non-negative random x)

$$\mathbf{E}[x] = \int_0^{\infty} x f(x) dx = \int_0^a x f(x) dx + \int_a^{\infty} x f(x) dx \geq \int_a^{\infty} a f(x) dx \geq a P(x \geq a)$$

Corollary: $P\left[|x| \geq \frac{1}{\sqrt{p}} \sqrt{E[x^2]}\right] \leq p$

$$P[x^2 \geq jE[x^2]] \leq \frac{1}{j} \Rightarrow P\left[|x| \geq \sqrt{jE[x^2]}\right] \leq \frac{1}{j}$$

$$R_i^T \leq \sum_{I \in \mathfrak{I}_i} R_i^{T,+}(I) = \sum_{I \in \mathfrak{I}_i} (R_i^{T,+}(I) - \tilde{R}_i^{T,+}(I) + \tilde{R}_i^{T,+}(I))$$

$$\leq \left| \sum_{I \in \mathfrak{I}_i} (R_i^{T,+}(I) - \tilde{R}_i^{T,+}(I)) \right| + \sum_{I \in \mathfrak{I}_i} \tilde{R}_i^{T,+}(I)$$

$$\leq \frac{1}{\sqrt{p}} \sqrt{E \left[\left(\sum_{I \in \mathfrak{I}_i} (R_i^{T,+}(I) - \tilde{R}_i^{T,+}(I)) \right)^2 \right]} + \frac{\Delta |\mathfrak{I}_i| \sqrt{|A|T}}{\delta}$$

$$\leq \left(\frac{|\mathfrak{I}_i| \sqrt{2}}{\sqrt{p}} + |\mathfrak{I}_i| \right) \frac{1}{\delta} \Delta \sqrt{|A|T}$$

MCCFR – Average Strategy



We need to maintain the average strategy without visiting.

Correct method

Note the strategy does not change without visits

Store additional information for later updates

$$w(I, a) = \sum_{t \in t_{last}, \dots, T} \pi_i^t(I) \sigma_i^t(I, a)$$

propagate down once sampled

Stochastically-weighted averaging:

Application of importance sampling

Boost the average strategy update by $1 / \text{probability of sampling } h$

May have high variance

CFR Variants – OOS



Online Outcome Sampling (Lisy et al. 2015)

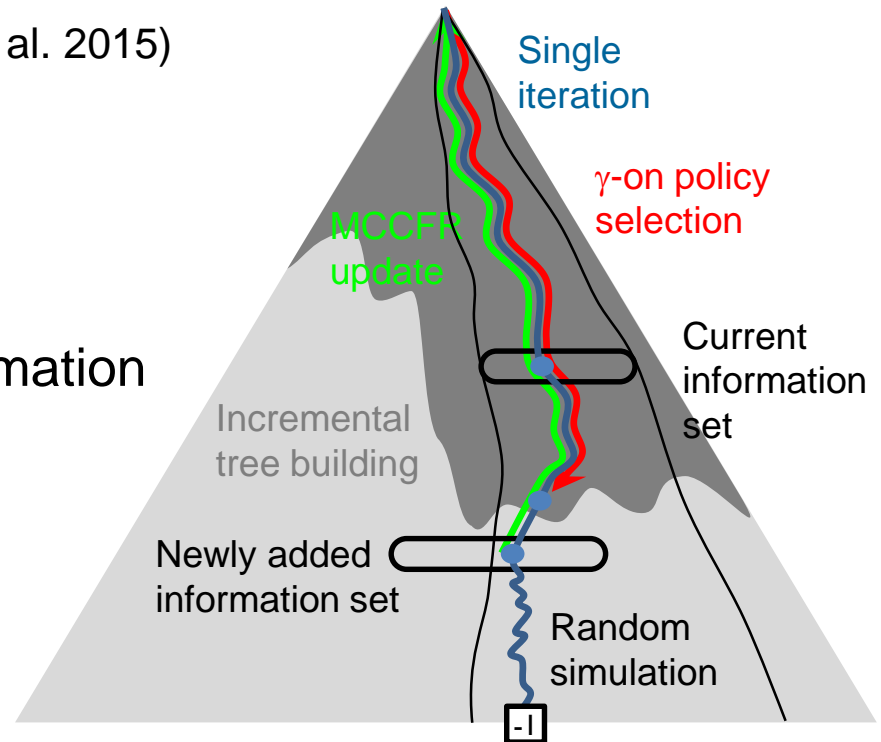
MCTS algorithm for imperfect information

Builds on MCCFR

Incremental tree building

Targets search to current IS

Guaranteed convergence to NE



CFR Variants – CFR-BR



Opponent always plays best response (Johanson et al. 2012)

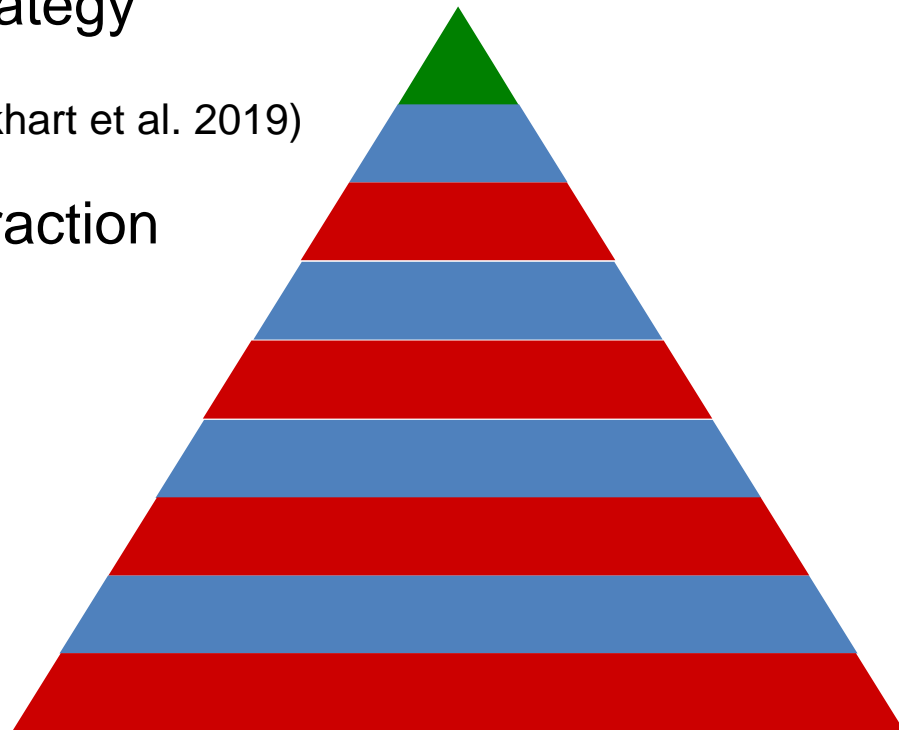
No storage for the opponent's strategy

No need for average strategy (Lockhart et al. 2019)

Opponent can play in a finer abstraction

Infinite strategy space

Optimal abstract strategies



CFR Variants – CFR-BR



Theorem (Johanson et al. 2012):

After T iterations, the average strategy of CFR-BR converges to

$$\frac{\Delta |I_1| \sqrt{|A_1|}}{\sqrt{T}} \text{- Nash equilibrium}$$

Proof sketch:

CFR player: $\sigma_i^0, \sigma_i^1, \dots, \sigma_i^T$ - no regret sequence of strategies

BR player: $BR(\sigma_i^0), BR(\sigma_i^1), \dots, BR(\sigma_i^T)$

Both players eventually have external regret $< \epsilon$

CFR Variants – CFR-BR



Theorem (Lockhart et al. 2019):

After T iterations, one of the **current** strategies produced by CFR-BR will be

$$\frac{\Delta |I_1| \sqrt{|A_1|}}{\sqrt{T}} \text{-Nash equilibrium}$$

Proof sketch:

- Any sequence of strategies with small cumulative regret includes an iterate with regret smaller than the average

Example CFR-BR + MCCFR



k-of-N robust optimization (Chen, Bowling 2012)

Optimal strategy for k worst samples from N

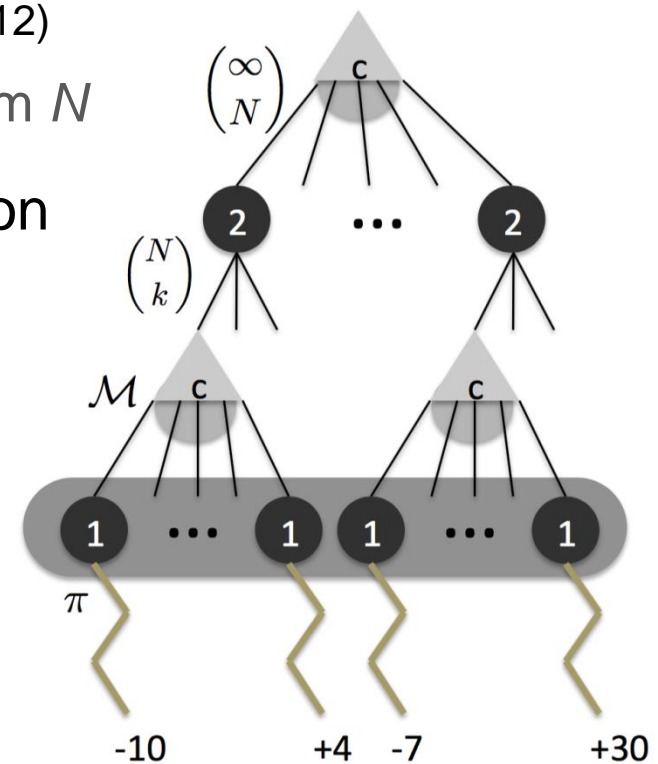
MDP with uncertainty in rewards/transition

Algorithm:

Sample a subgame (MCCFR)

Pick BR for player 2

Update player 1 using CFR



References



Zinkevich, M., Johanson, M., Bowling, M., & Piccione, C. (2008). Regret minimization in games with incomplete information. *Advances in Neural Information Processing Systems*, 20, 1729–1736.

Lanctot, M. (2013.). Monte Carlo Sampling and Regret Minimization for Equilibrium Computation and Decision-Making in Large Extensive Form Games. PhD Thesis. University of Alberta.

Lockhart, E., Lanctot, M., Perolat, J., Lespiau, J., Morrill, D., Timbers, F., Tuyls, K. (2019). Computing Approximate Equilibria in Sequential Adversarial Games by Exploitability Descent. arXiv:1903.05614v2.

Chen, K., & Bowling, M. (2012). Tractable Objectives for Robust Policy Optimization. *Advances in Neural Information Processing Systems 25 (NIPS)*, 2078-2086.