Algorithmic Game Theory

Learning in Games

Viliam Lisý

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

(May 18, 2018)

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

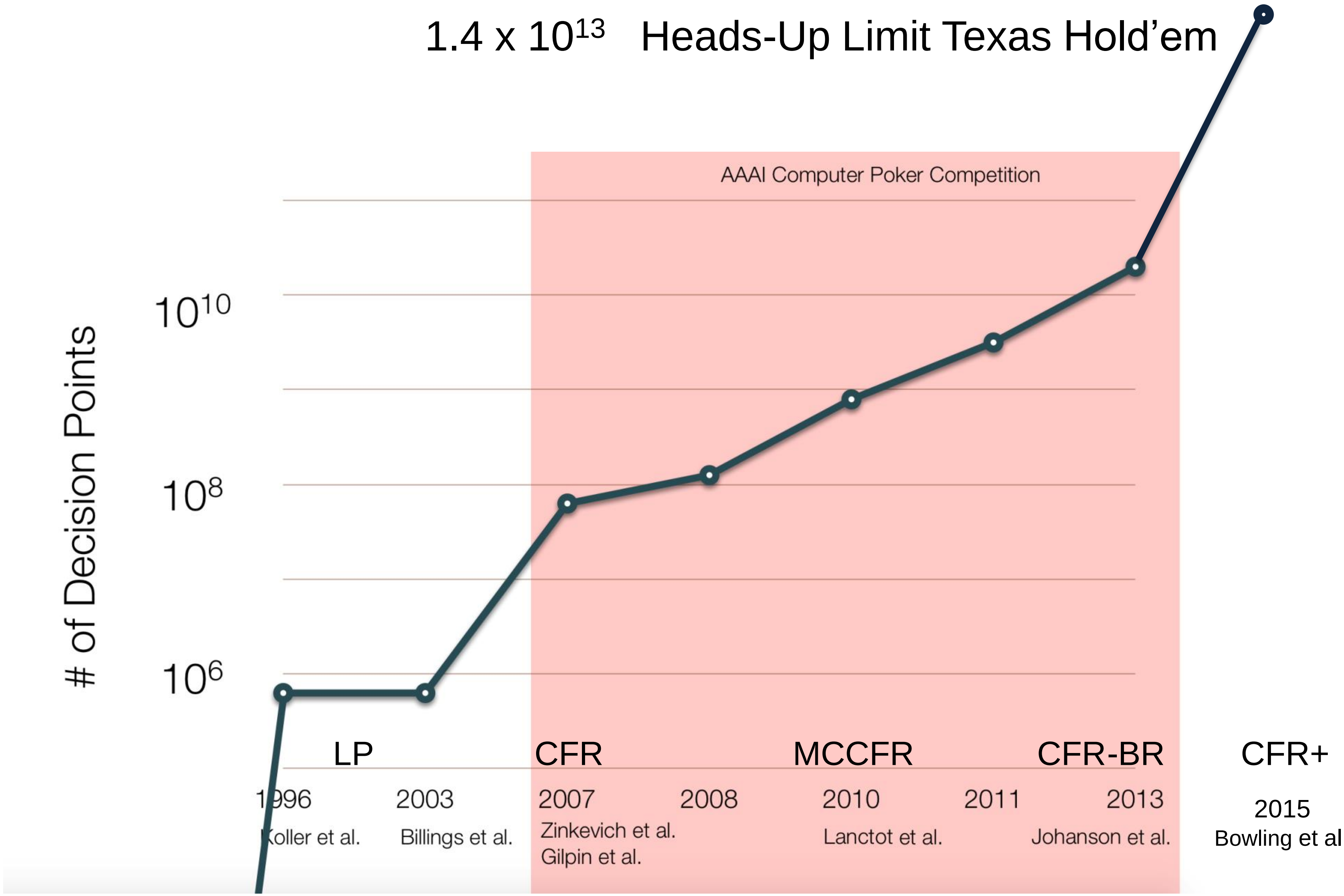
Algorithmic Game Theory
Learning in extensive form games

Viliam Lisý

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

(May 15, 2017)

Impact on poker performance



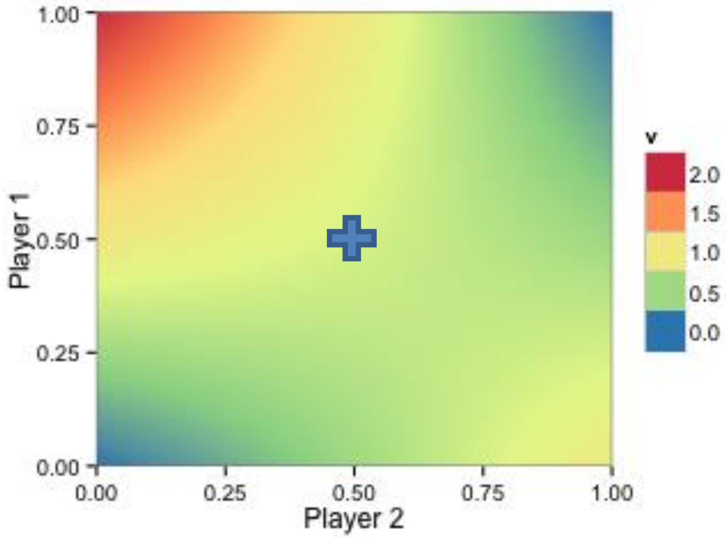
Based on M. Bowling's slide from AAAI 2015 keynote

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+

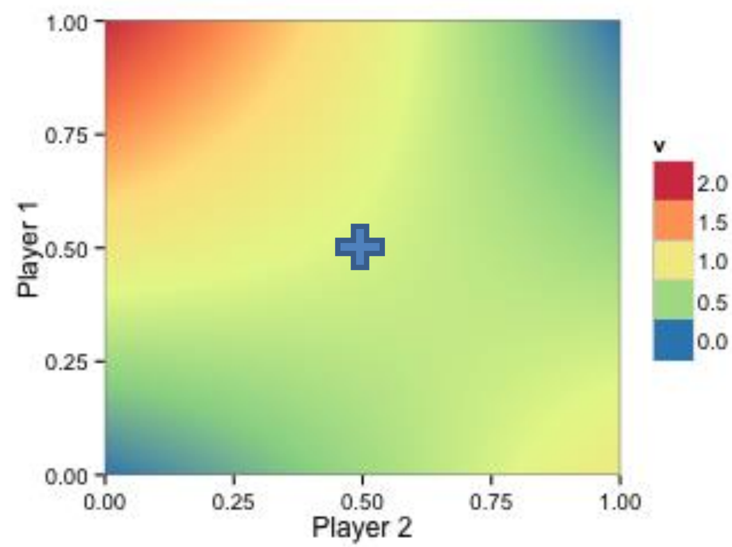
σ^t	0.5	0.5
0.5	2	0
0.5	0	1



Regret matching+

Iteration:
1

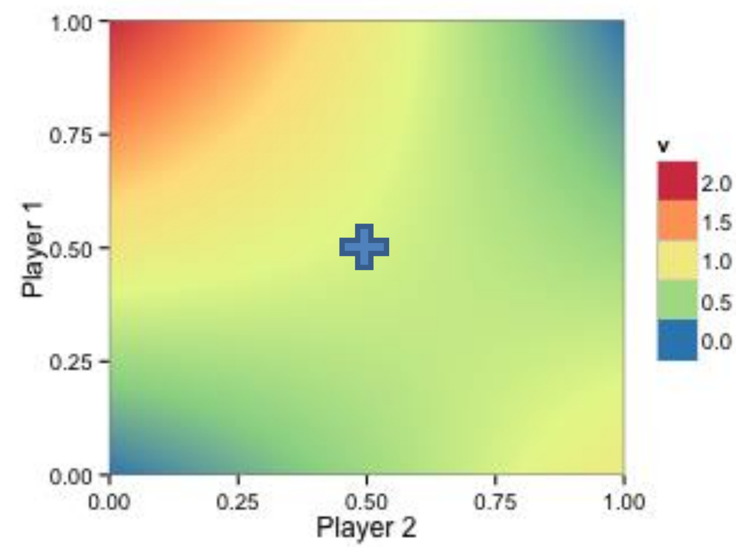
			$\overline{\sigma_2}$		
			R_2	0	0
			r_2		
$\overline{\sigma_1}$	R_1	r_1	σ^t	0.5	0.5
	0	0.25	0.5	2	0
	0	-0.25	0.5	0	1



Regret matching+

Iteration:
1

			$\overline{\sigma_2}$		
			R_2	0	0
			r_2		
$\overline{\sigma_1}$	R_1	r_1	σ^t	0.5	0.5
	0.25	0.25	0.5	2	0
	0	-0.25	0.5	0	1



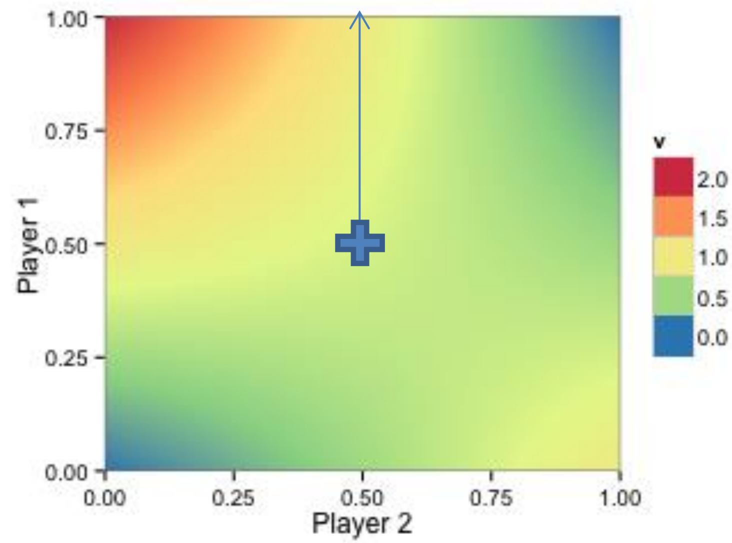
Regret matching+

Iteration:
1

$\overline{\sigma_2}$		
R_2	0	0
r_2		
σ^t	0.5	0.5

$\overline{\sigma_1}$	R_1	r_1
	0.25	0.25
	0	-0.25

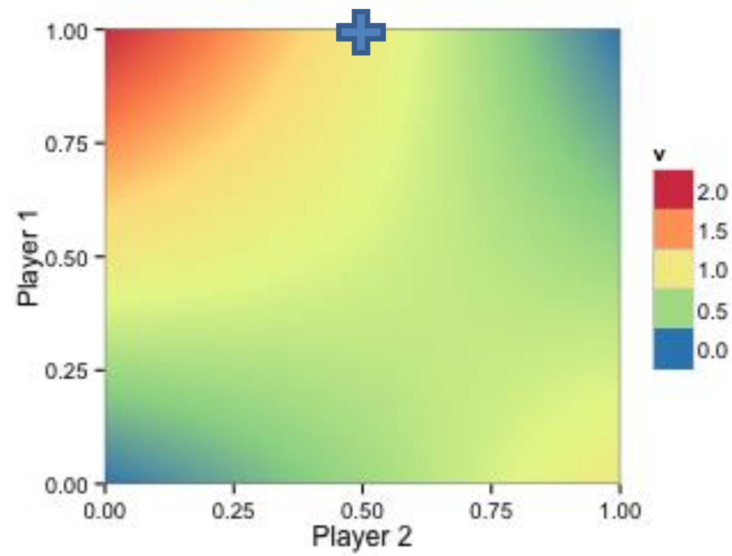
2	0
0	1



Regret matching+

Iteration:
1

			$\overline{\sigma_2}$		
			R_2	0	0
			r_2	-1	1
$\overline{\sigma_1}$	R_1	r_1	σ^t	0.5	0.5
1	0.25		1	2	0
0	0		0	0	1

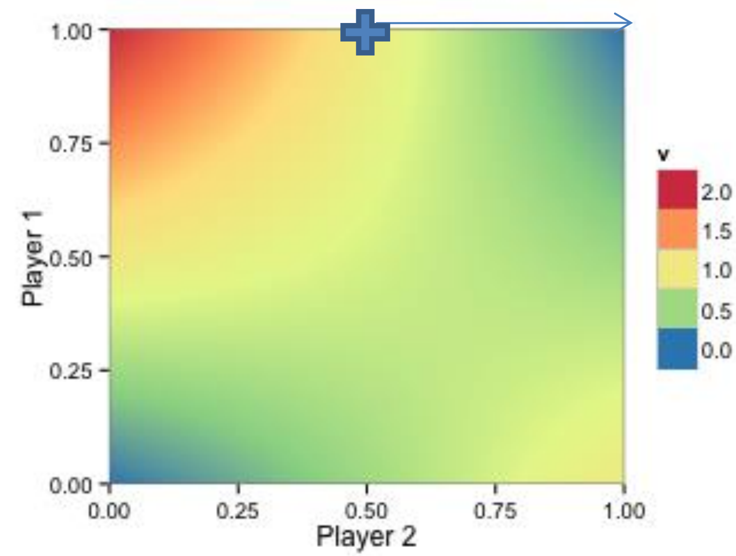


Regret matching+

Iteration:
1

$\overline{\sigma_2}$		
R_2	0	1
r_2	-1	1

$\overline{\sigma_1}$	R_1	r_1	σ^t		
1	0.25		1	2	0
0	0		0	0	1

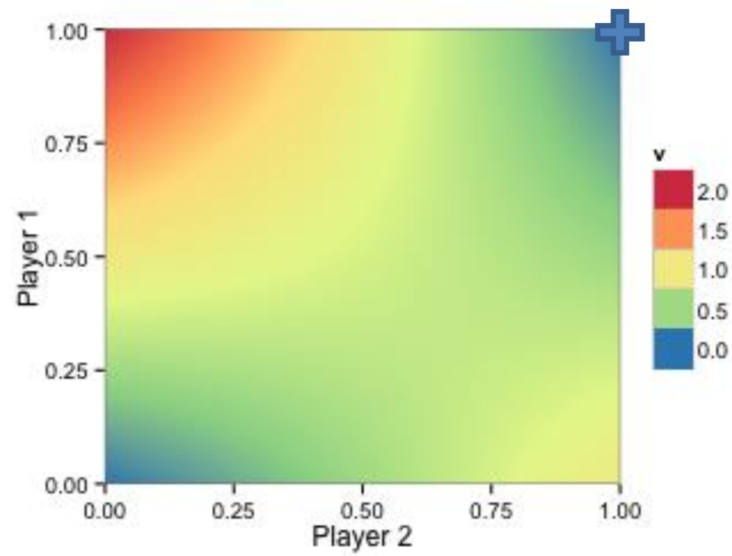


Regret matching+

Iteration:
2

$\overline{\sigma_2}$	0	1
R_2	0	1
r_2		

$\overline{\sigma_1}$	R_1	r_1	σ^t	0	1
1	0.25	0	1	2	0
0	0	1	0	0	1



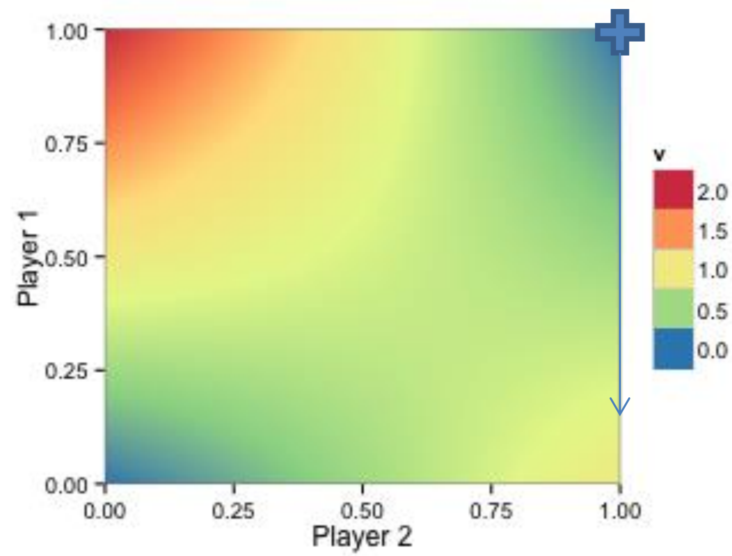
Regret matching+

Iteration:
2

$\overline{\sigma_2}$	0	1
R_2	0	1
r_2		
σ^t	0	1

$\overline{\sigma_1}$	R_1	r_1
1	0.25	0
0	1	1

2	0
0	1

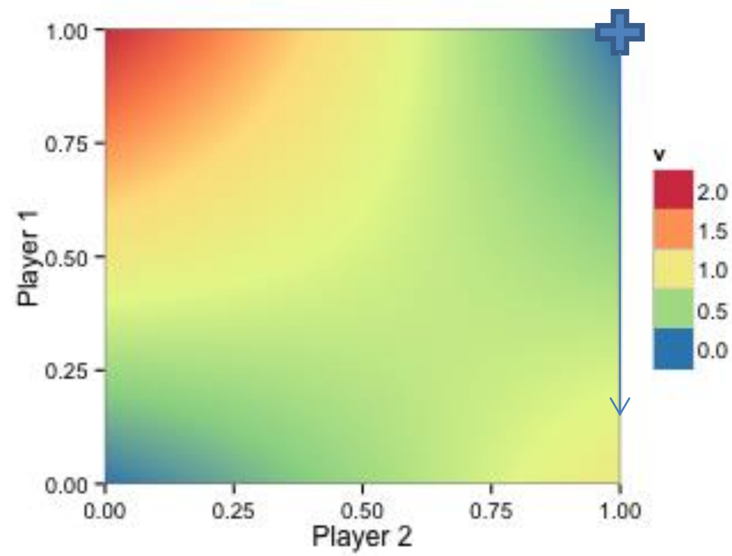


Regret matching+

Iteration:
2

$\overline{\sigma_2}$	0	1
R_2	0	1
r_2		
σ^t	0	1

$\overline{\sigma_1}$	R_1	r_1	σ^t		
0.46	0.25	0	0.2	2	0
0.54	1	1	0.8	0	1

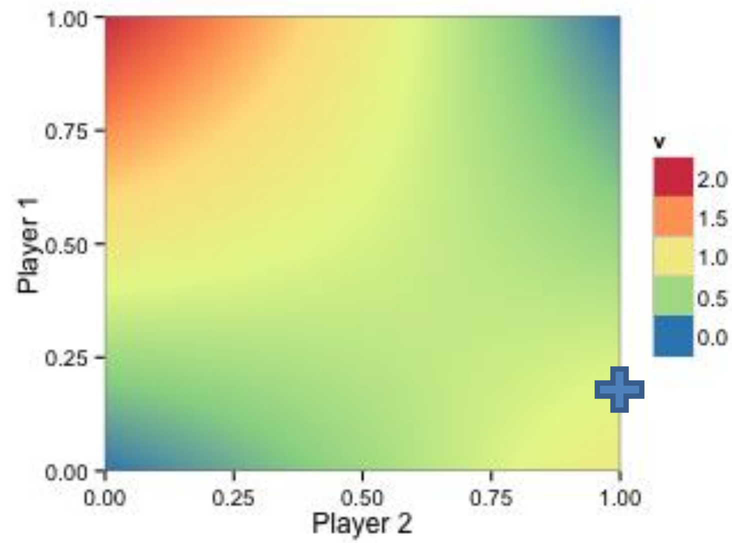


Regret matching+

Iteration:
2

$\overline{\sigma_2}$		0	1
R_2		0	1
r_2		0.4	0
σ^t		0	1

$\overline{\sigma_1}$	R_1	r_1	σ^t		
0.46	0.25		0.2	2	0
0.54	1		0.8	0	1

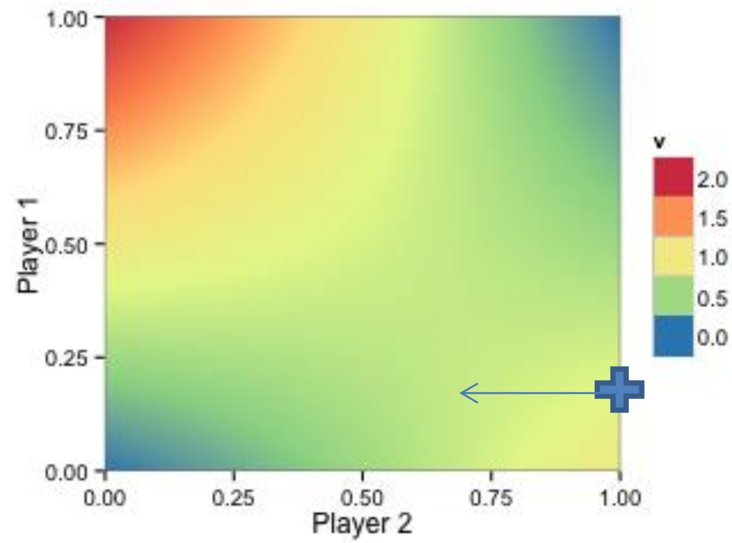


Regret matching+

Iteration:
2

$\overline{\sigma_2}$	0	1
R_2	0.4	1
r_2	0.4	0

$\overline{\sigma_1}$	R_1	r_1	σ^t		
0.46	0.25		0.2	2	0
0.54	1		0.8	0	1

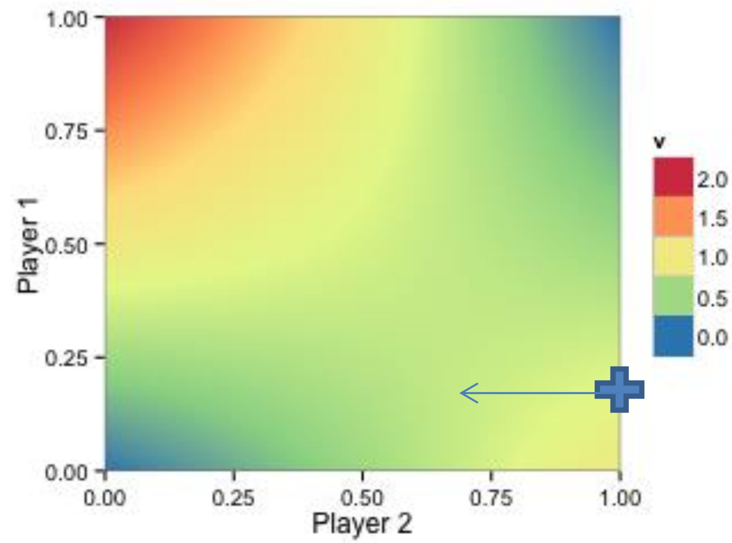


Regret matching+

Iteration:
2

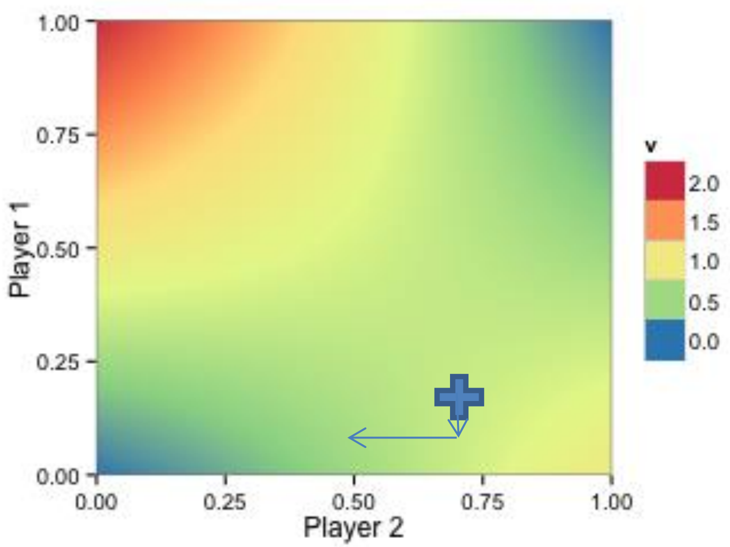
$\overline{\sigma_2}$	0.19	0.81
R_2	0.4	1
r_2	0.4	0
σ^t	0.29	0.71

$\overline{\sigma_1}$	R_1	r_1	σ^t		
0.46	0.25		0.2	2	0
0.54	1		0.8	0	1



Regret matching+

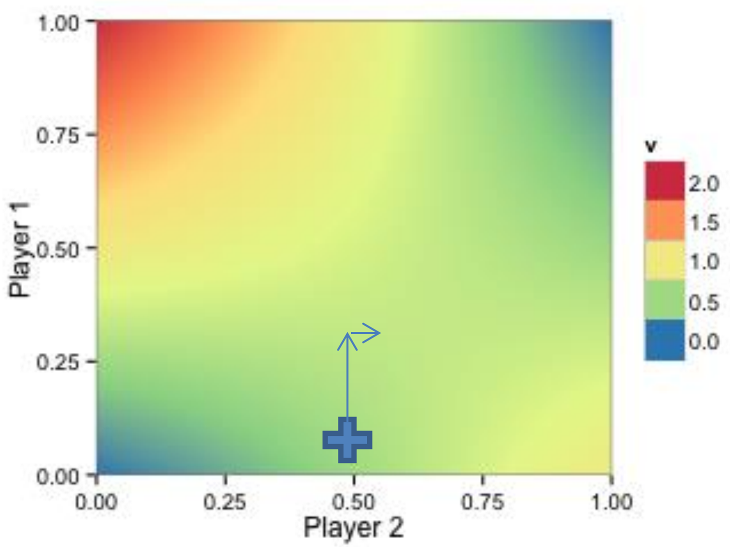
Iteration:
3



Regret matching+

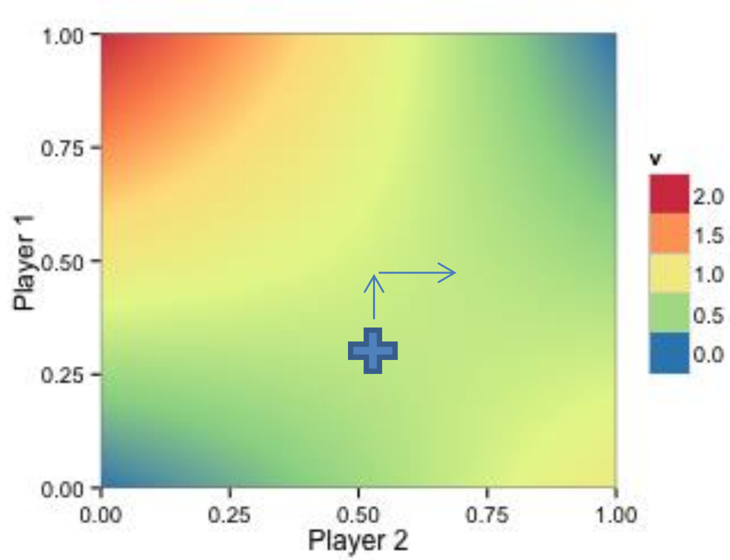
Iteration:

4



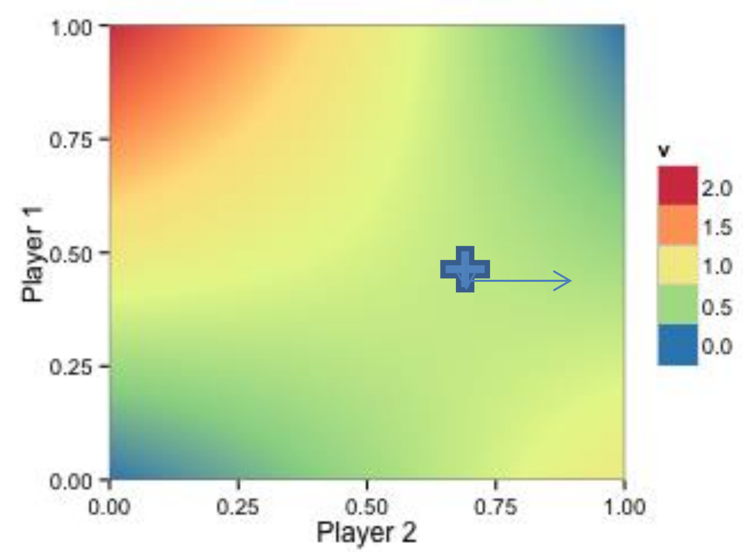
Regret matching+

Iteration:
5



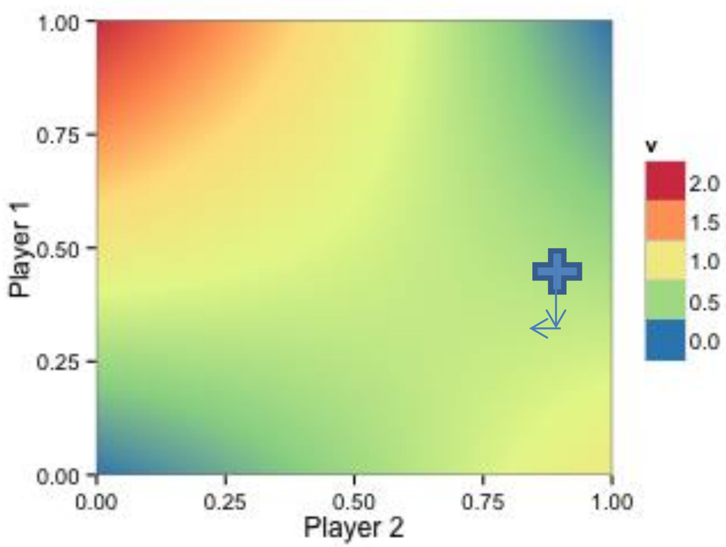
Regret matching+

Iteration:
6



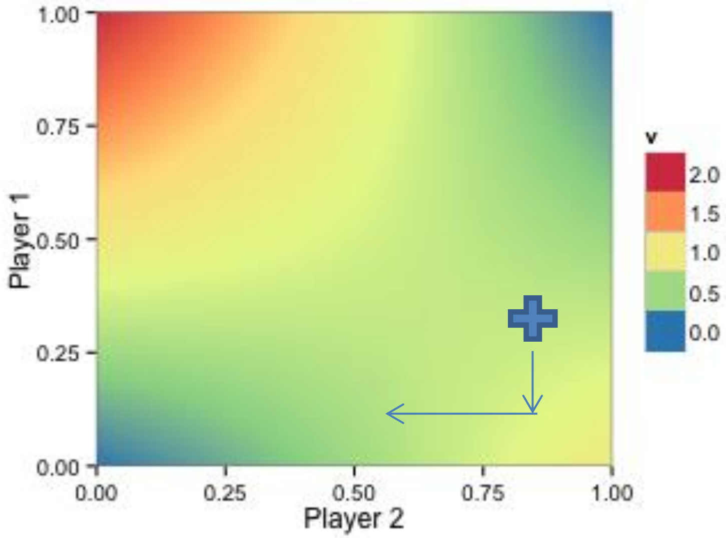
Regret matching+

Iteration:
7



Regret matching+

Iteration:
8

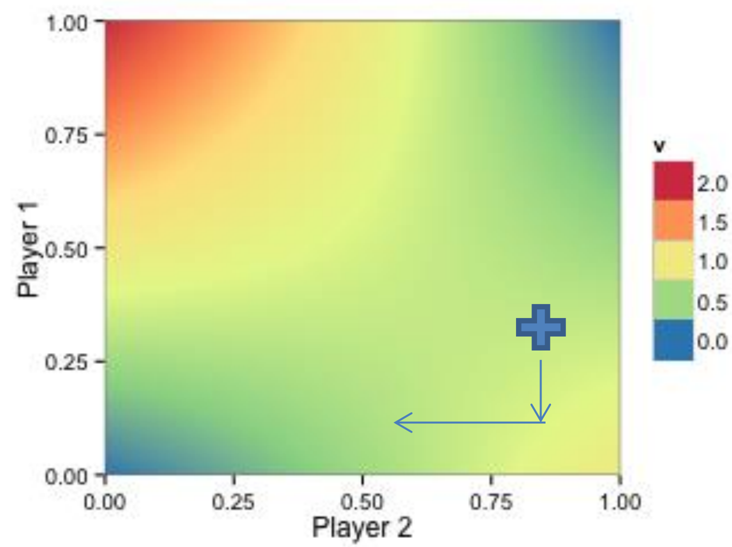


Regret matching+

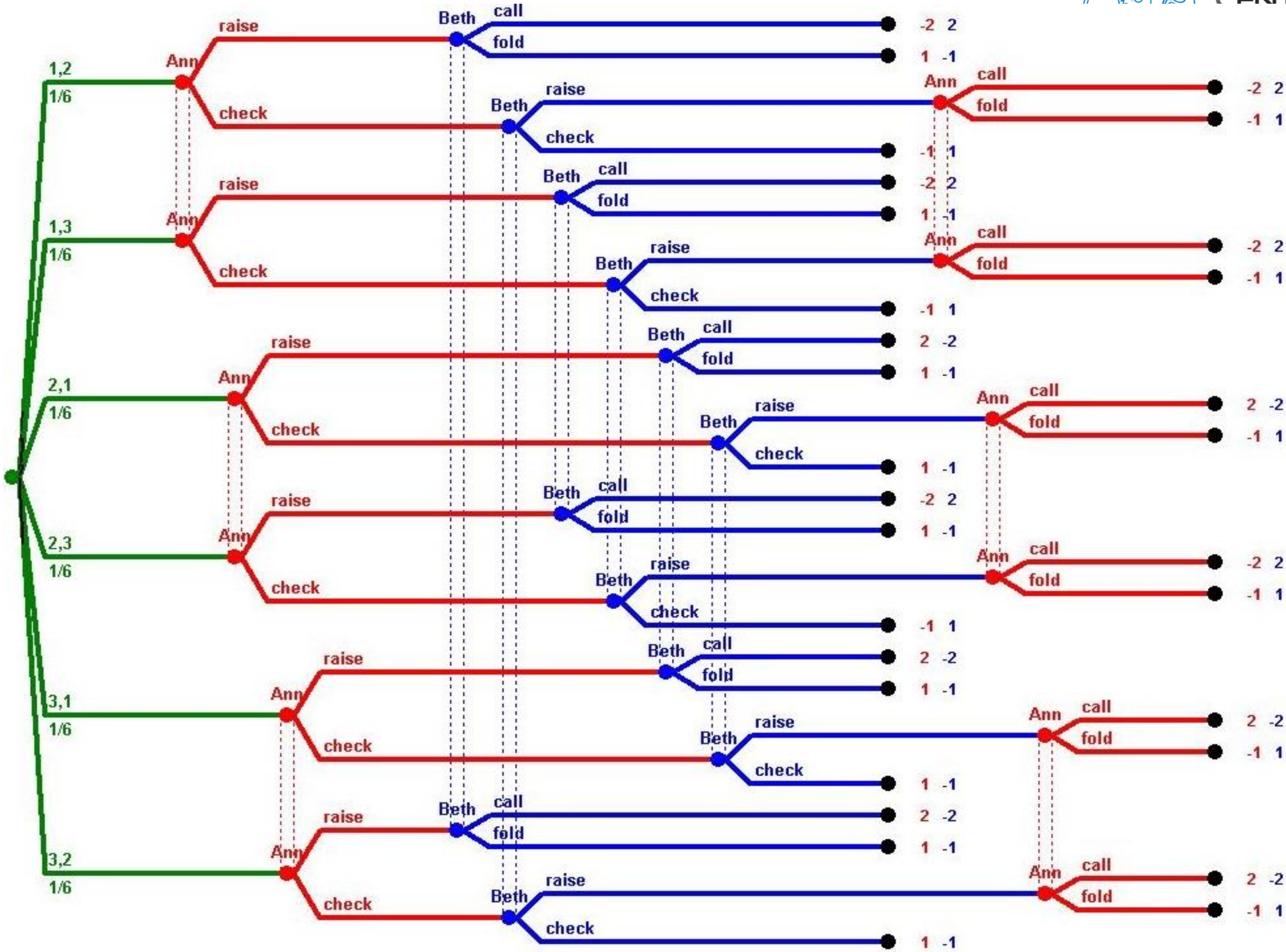
Iteration:
8

$\overline{\sigma_2}$	0.30	0.70
R_2	0.83	1.15
r_2		

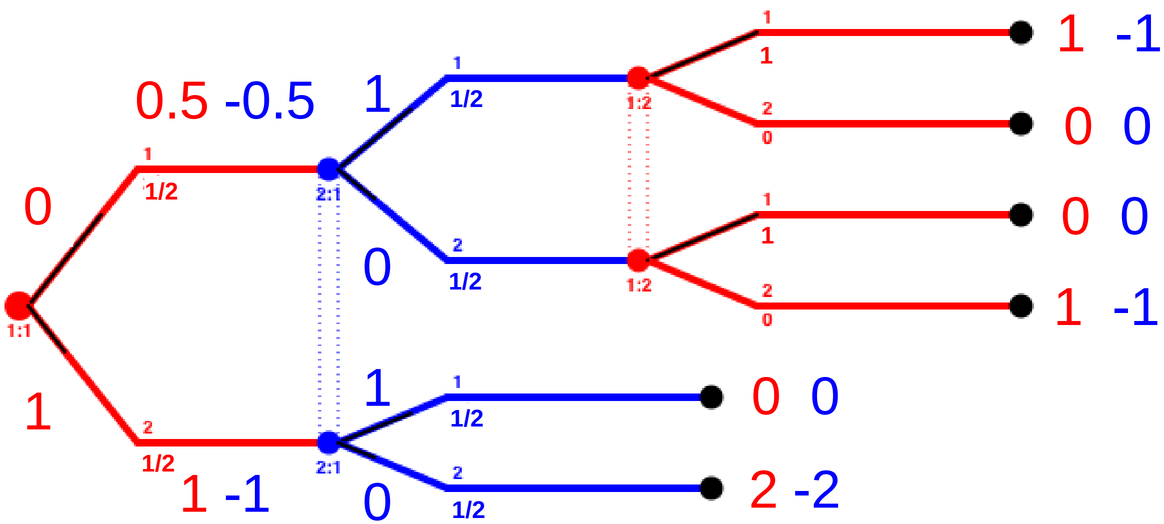
$\overline{\sigma_1}$	R_1	r_1	σ^t	0.42	0.58
0.33	0.17		0.11	2	0
0.67	1.30		0.88	0	1



Extensive form games



Counterfactual Regret - Motivation



1	0
0	1

0	2

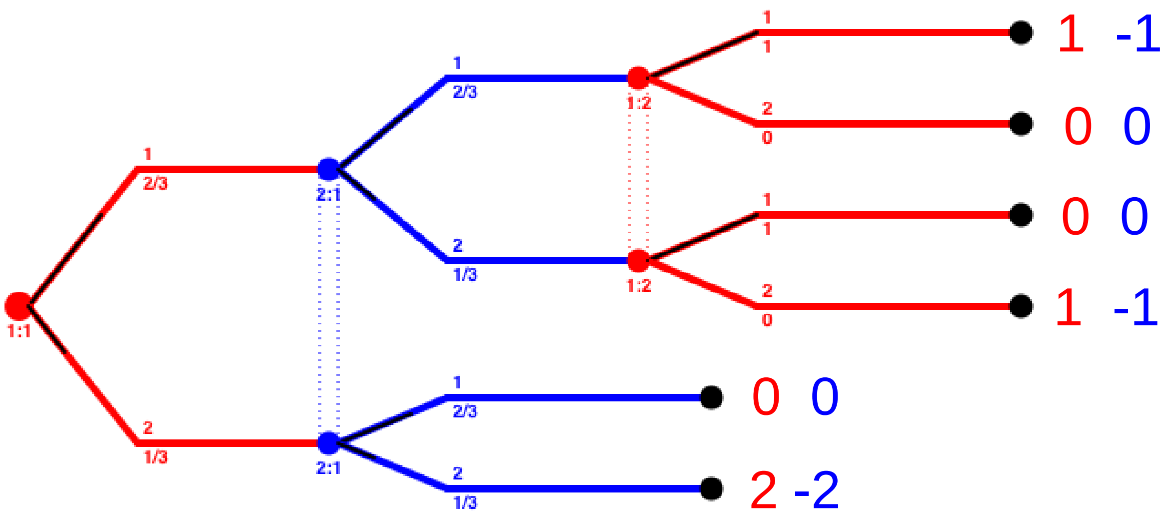
Take the current reach probabilities?

-> 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 conterfactual 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

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, $Succ_i(I) = \bigcup_{a\in A(I)} Succ_i(I,a)$.

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

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

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

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

$$R_{i,full}^{T}(I) = \max_{a \in A(I)} \max_{\sigma' \in \Sigma_i} \sum_{t \in 1..T}$$

$$(v_i(\sigma^t|_{I \to a}, I) - v_i(\sigma^t, I)$$

$$+ \sum_{I' \in Succ_i(I,a)} succ_i^{\sigma}(I'|I,a)(\frac{\pi_{-i}^{\sigma^t}(I)}{\pi_{-i}^{\sigma^t}(I')})\left(v_i\left(\sigma^t|_{D(I) \to \sigma'}, I'\right) - v_i(\sigma^t, I')\right))$$

$$R_{i,full}^{T}(I) \leq \max_{a \in A(I)} \max_{\sigma' \in \Sigma_i} \sum_{t \in 1..T} (v_i(\sigma^t|_{I \to a}, I) - v_i(\sigma^t, I))$$

$$+ \max_{a \in A(I)} \max_{\sigma' \in \Sigma_i} \sum_{t \in 1..T} \sum_{I' \in Succ_i(I,a)} \left(v_i\left(\sigma^t|_{D(I') \to \sigma'}, I'\right) - v_i(\sigma^t, I')\right)$$

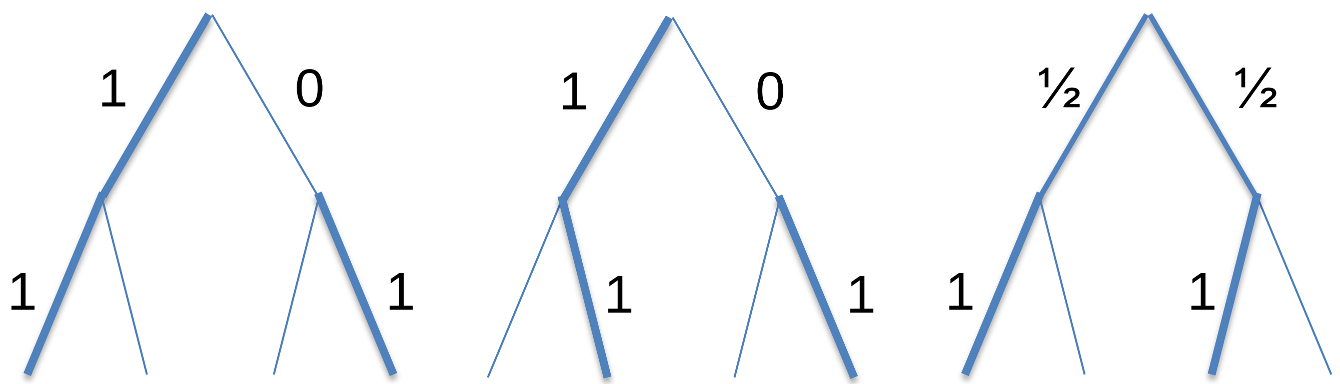
$$R_{i,full}^{T}(I) \leq R_{i,imm}^{T}(I) + \max_{a \in A(I)} \sum_{I' \in Succ_i(I,a)} R_{i,full}^{T}(I')$$

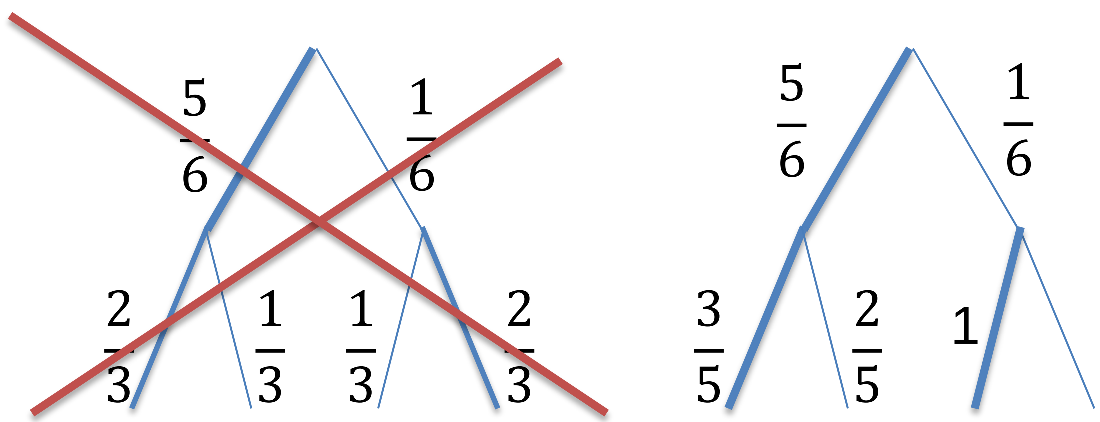
$$\leq R_{i,imm}^{T}(I) + \sum_{I' \in Succ_i(I)} R_{i,full}^{T,+}(I').$$

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

Average Strategy in CFR

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

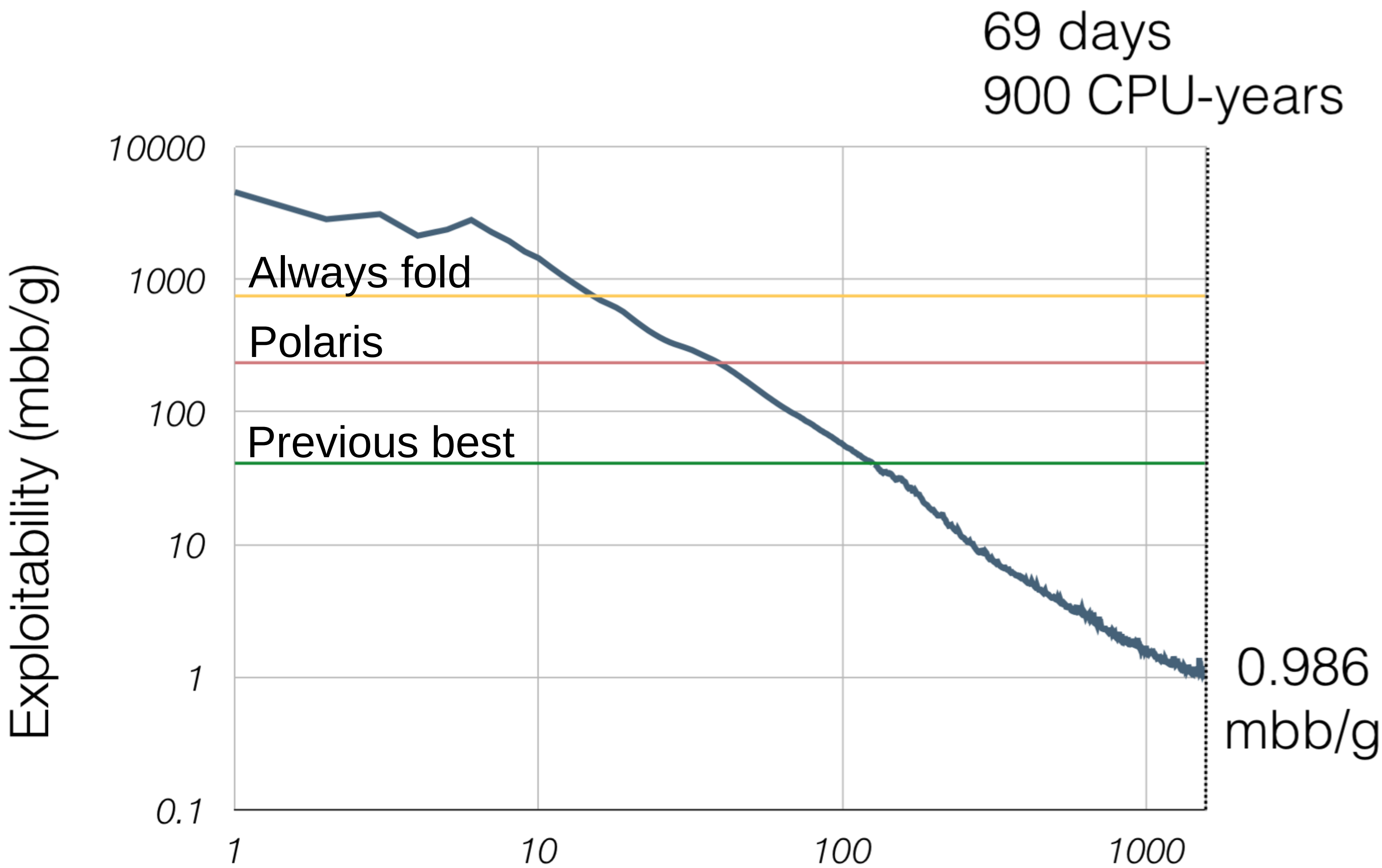




CFR+ Convergence Speed

Theorem (Tammelin et al. 2015): The mean strategies form 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., Sience 2015)



Slide from M. Bowling's AAAI 2015 keynote

CFR Variants – MCCFR

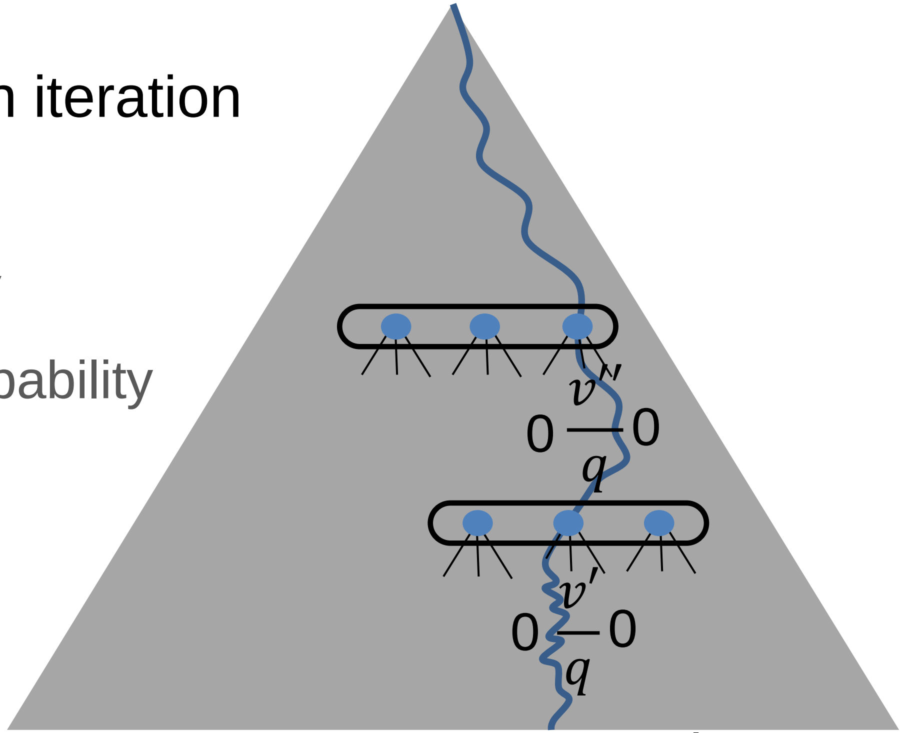
Monte Carlo Counterfactural Regret Minimization (Lanctot et al. 2009)

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



$\frac{v}{q}$ - value, - probability

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 $Q_j \in Q$,

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

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

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

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

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

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

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

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

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

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
- $w(I,a) = \sum_{t \in t_{last},..,T} \pi_i^t(I) \sigma_i^t(I,a)$
- propagate down once sampled

Stochastically-weighted averaging:

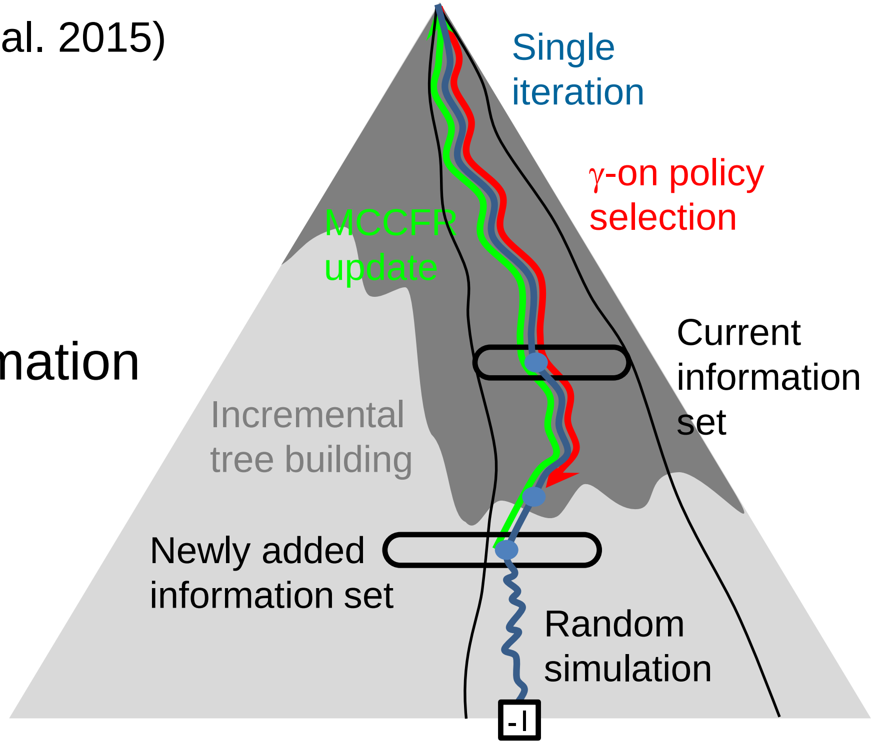
- Application of importance sampling
- Boost the average strategy update by 1 / probability of sampling h
- May have high variance

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

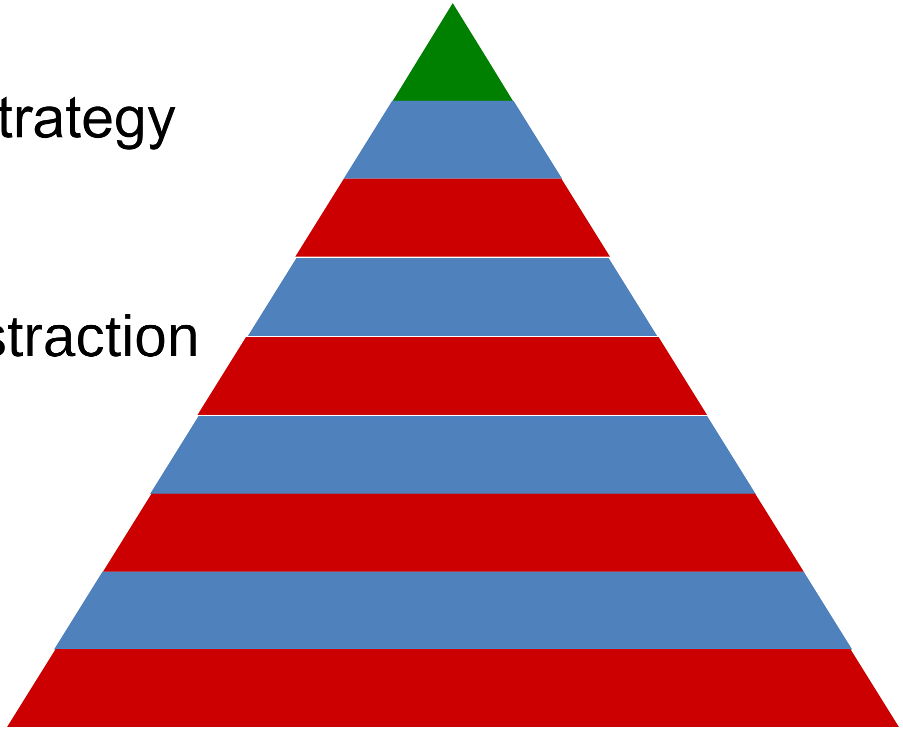
Opponent always plays best response (Johanson et al. 2012)

No storage for the opponent's strategy

No need for average strategy

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

$$\frac{\Delta|\mathrm{I}_1|\sqrt{|A_1|}}{\sqrt{T}}\text{-Nash equilibrium}$$

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 (Johanson et al. 2012):

After T iteration with probability (1-p) the **current strategy** of CFR-BR converges to

$\frac{\Delta|\mathrm{I}_1|\sqrt{|A_1|}}{p\sqrt{T}}$-Nash equilibrium

Proof sketch:

$$\bar{r}^T_{i,full} = \frac{1}{T}\max_{\sigma'}\sum_{t=1}^{T} u_i(\sigma', \sigma^t_{-i}) - \frac{1}{T}\sum_{t=1}^{T} u_i\left(\sigma^t_i, \sigma^t_{-i}\right) < \epsilon$$

$$\frac{1}{T}\sum_{t=1}^{T} u_i\left(\sigma^t_i, \sigma^t_{-i}\right) \geq \frac{1}{T}\max_{\sigma'}\sum_{t=1}^{T} u_i\left(\sigma', \sigma^t_{-i}\right) - \epsilon \geq \max_{\sigma'} u_i\left(\sigma', \bar{\sigma}^T_{-i}\right) - \epsilon$$

$\geq v^*_i - \epsilon$, but $u_i\left(\sigma^t_i, \sigma^t_{-i}\right) \leq v^*_i$, therefore $u_i\left(\sigma^t_i, \sigma^t_{-i}\right) > v^*_i - \frac{\epsilon}{p}$ often.

Example CFR-BR + MCCFR

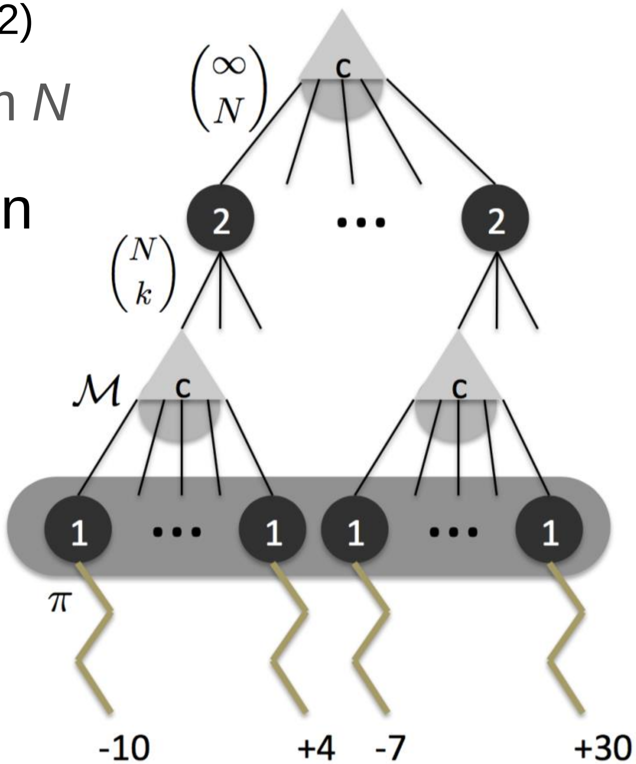
k-of-N robust optimization (Chen, Bowling 2012)

- 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.

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