Approaching the Nash Equilibrium

An Introduction to Counterfactual Regret and Improvements

Michael Schubert

TNG Technology Consulting GmbH, www.tngtech.com

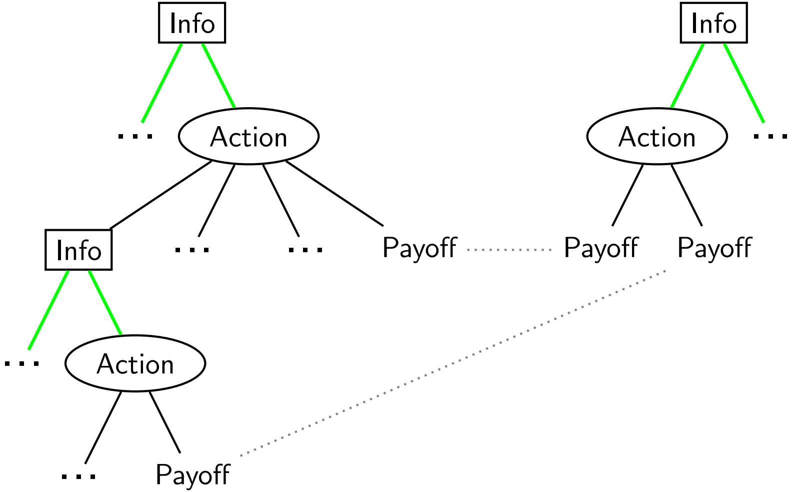
Big Tech Day 10, 2.6.2017

Outline

Basic Notions

- Information point I with set of possible actions $A(I)$. Every action a belongs to a unique info set.
- Action leads to new information point in $\mathbf{I}(a)$ or payoff in $P(A)$, depending on chance and opponents actions.
- Full game history h: Sequence of actions $A_1(h)$ and $A_2(h)$, and chance. Corresponding to a payoff.

Game Tree


Info
Action
...
Info
...
...
Payoff
Action
...
...
Payoff
Info
Action
...
Payoff
Payoff

Worked Examples

- Matrix Games: Player 1 chooses row, Player 2 chooses column, payoff is the corresponding entry.
- Leduc Hold'em:
  - Three types of cards, two of cards of each type.
  - Betting round - Flop - Betting round.
  - Fixed betting amount per round (e.g. 2 and 4), at most one bet and one raise.
  - Player with same card as flop wins, else highest card.
- Leduc-5: Same as Leduc, just with five different betting amounts (e.g. 1, 2, 4, 8, 16 and twice as much in round 2) per round.

Strategies and Expectation

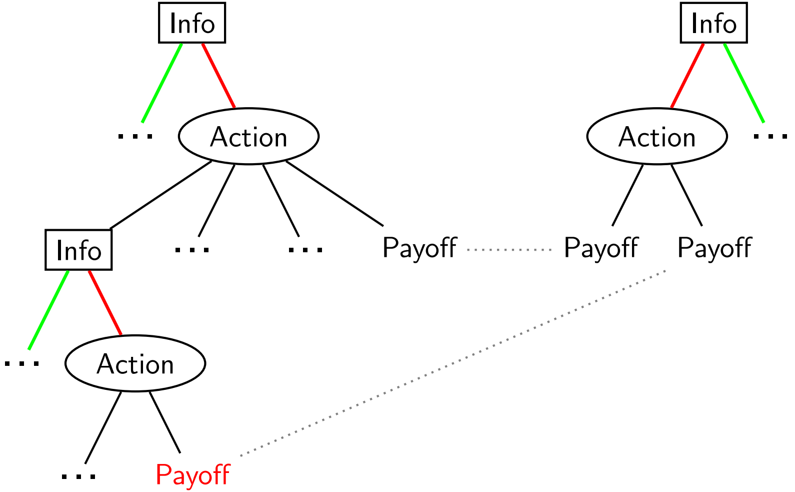
- Strategy: Assigns probability $\sigma(a)$ to players actions.
- Expectation:

$$E_i(\sigma) = \sum_{h \in H} \prod_{a \in A_1(h)} \sigma(a) \prod_{a \in A_2(h)} \sigma(a) \cdot v_i(h)$$

- Two Player-Zero Sum Games:

$$v_1(h) = -v_2(h)$$

Calculating Expectation



Exploitability

Calculate best response σ^1_{BR} to σ^2 and σ^2_{BR} to σ^1.

$$\text{expl}(\sigma) = E_1\left(\sigma^1_{\text{BR}}, \sigma^2\right) + E_2\left(\sigma^1, \sigma^2_{\text{BR}}\right)$$

Nash Equilibrium: $\text{expl}(\sigma) = 0$

LP-Form

$$\tilde{\sigma}(a) = \pi^i_\sigma(I) \cdot \sigma(a)$$

Constraints:

$$\forall I \in \mathbf{I}(a_0) : \sum_{a \in A(I)} \tilde{\sigma}(a) = \tilde{\sigma}(a_0)$$

If I is an entry point:

$$\sum_{a \in A(I)} \tilde{\sigma}(a) = 1$$

$\rightarrow$ Strategy Polyhedron. Expectation is the bilinear form:

$$E_i(\sigma) = \sum_{h \in H} \tilde{\sigma}\left(a^0_{\text{fin}}(h)\right) \cdot \tilde{\sigma}\left(a^1_{\text{fin}}(h)\right) \cdot v_i(h)$$

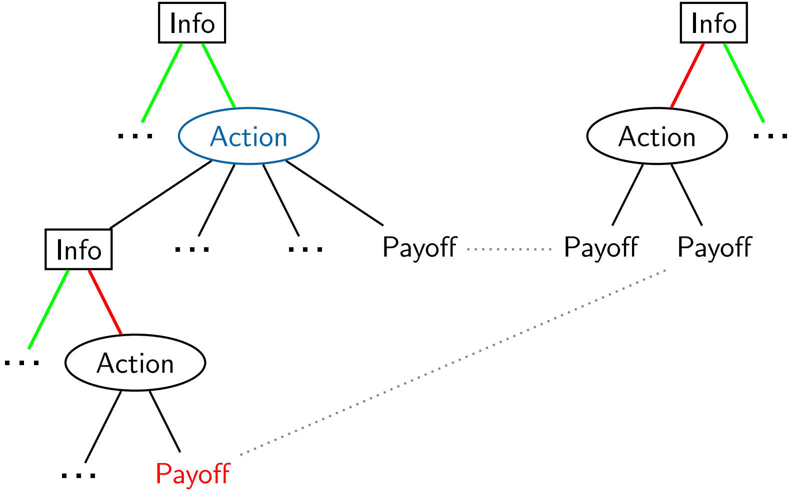
CFR: Basic Notions

Counterfactual value of information points and actions:

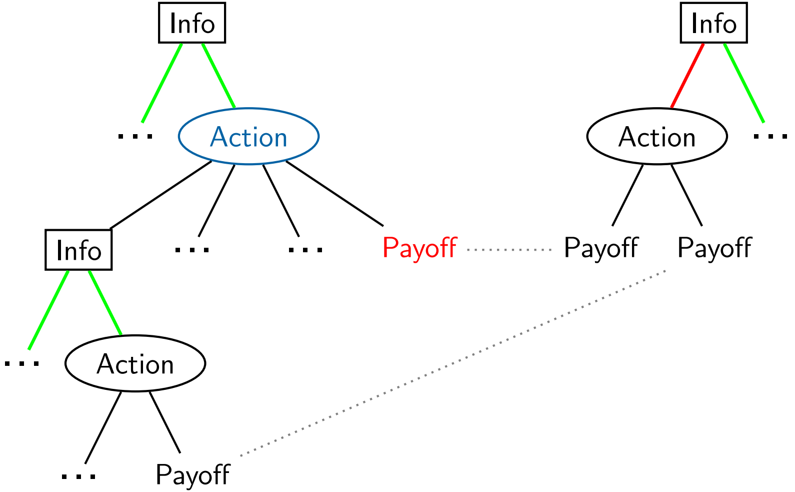
$$v_\sigma(I) = E_\sigma(I) \cdot \pi_\sigma^{-i}(I)$$

$$v_\sigma(a) = E_\sigma(a) \cdot \pi_\sigma^{-i}(I)$$

Calculating Counterfactual Value



Calculating Counterfactual Value



CFR: Basic Notions

Counterfactual Value:

$$v_\sigma(I) = E_\sigma(I) \cdot \pi_\sigma^{-i}(I)$$

$$v_\sigma(a) = E_\sigma(a) \cdot \pi_\sigma^{-i}(I)$$

Immediate Counterfactual Regret:

$$r_\sigma(a) = v_\sigma(a) - v_\sigma(I)$$

Counterfactual Best Response: Maximizes all counterfactual values for a fixed opponents strategy.

Regret Matching and CFR+

Cumulative counterfactual regret:

$$R_t(a) = R_{t-1}(a) + r_{\sigma_t}(a)$$

Regret matching

$$\sigma_{t+1}(a) = \frac{R_t(a)^+}{\sum_{a \in I} R_t(a)^+}$$

CFR+:

$$R_t^+(a) = \max\left(R_{t-1}^+(a) + r_{\sigma_t}(a), 0\right)$$

Averaging

Average strategy for action a by player i following information I:

$$\bar{\sigma}_T(a) = \frac{\sum_{t=0}^{T} \pi^{i}_{\sigma_t}(I)\sigma_t(a)}{\sum_{t=0}^{T} \pi_{\sigma_t}(I)}$$

Leads to

$$\bar{\tilde{\sigma}}_T(a) = \frac{\sum_{t=0}^{T} \tilde{\sigma}_t(a)}{T+1}$$

Weighted Averaging

Average strategy:

$$\bar{\sigma}_T(a) = \frac{\sum_{t=0}^{T} w_t \pi_{\sigma_t}(I)\sigma_t(a)}{\sum_{t=0}^{T} w_t \pi_{\sigma_t}(I)}$$

Leads to

$$\bar{\tilde{\sigma}}_T(a) = \frac{\sum_{t=0}^{T} w_t \tilde{\sigma}_t(a)}{\sum_{t=0}^{T} w_t}$$

Common weights: $w_t = t$ (linear averaging) or $w_t = t^2$ (squared averaging).

Observations

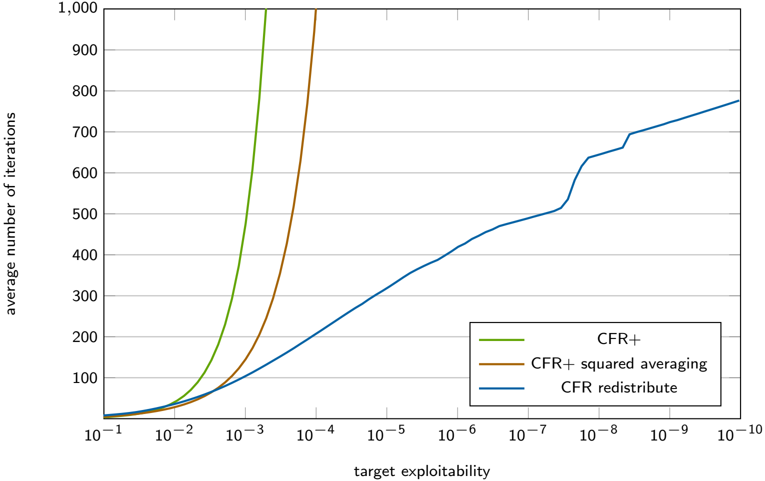
- Average Strategy has much less exploitability than current strategy.
- CFR+ current strategy has signlificantly less exploitability than CFR current strategy.
- Big impact of weighted averaging with CFR+: The average strategy is more likely to be close to the Nash Equilibrium when the current strategies are close to the Nash Equilibrium as well.

Idea: Continue CFR-iterations at average strategy, by matching regret with the average strategy after a number of Iterations:

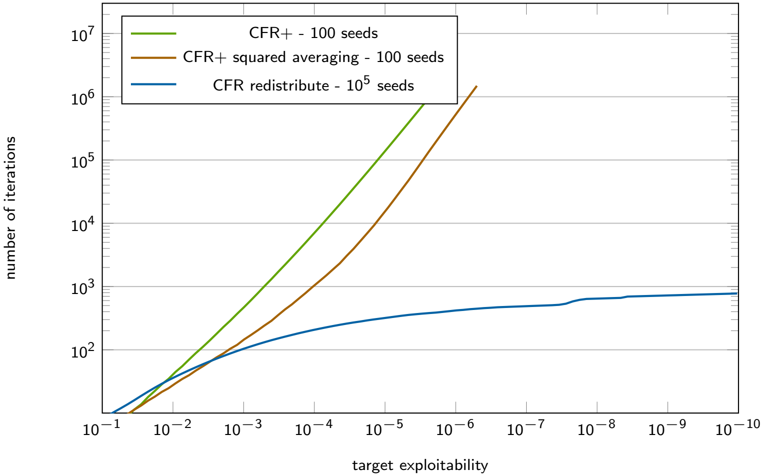
$$R_{\text{new}}(a) = \bar{\sigma}(a) \cdot \sum_{a' \in A(I)} R_{\text{old}}(a')^{+}$$

Empirical good blocklength: $n \log(n)$ for the n^{th} block.

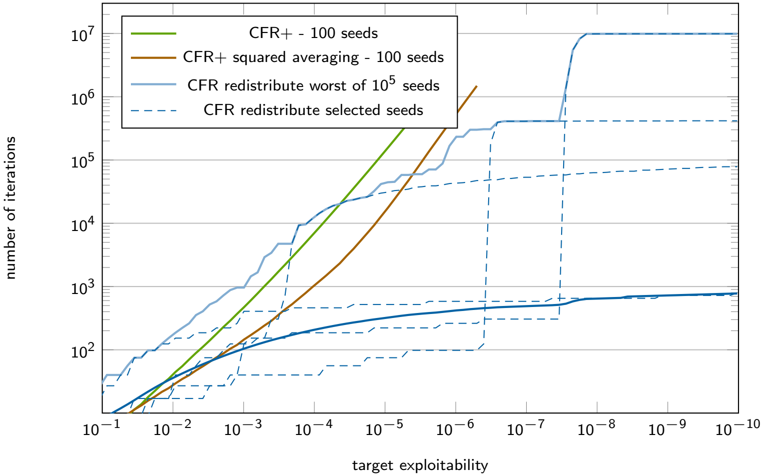
Random 10×10 - Matrix Game (10^5 Seeds)



Random 10 × 10 - Matrix Game



Random 10 × 10 - Matrix Game



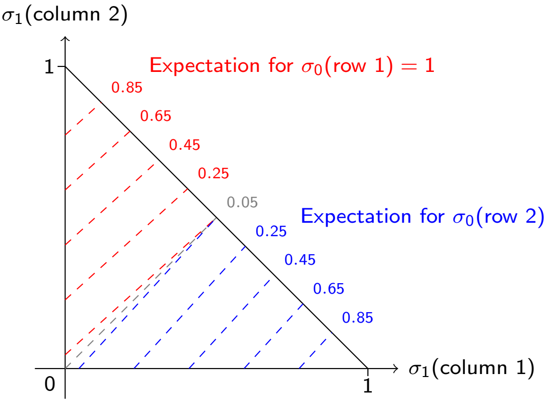
Problems

- Mainly heuristic argumentation - No guarantees for convergence
- Noisy performance
- Sometimes adjusting the quantity of the regret is needed
- Very different behaviour for similar games

- Exploitability is not equivalent to distance from Nash equilibrium. Example: Matrix game with payoff

$$P = \begin{pmatrix} 1+\epsilon & -1 & 0 \\ -1 & 1+\epsilon & 0 \end{pmatrix}$$

$$P = \begin{pmatrix} 1+10^{-1} & -1 & 0 \\ -1 & 1+10^{-1} & 0 \end{pmatrix}$$



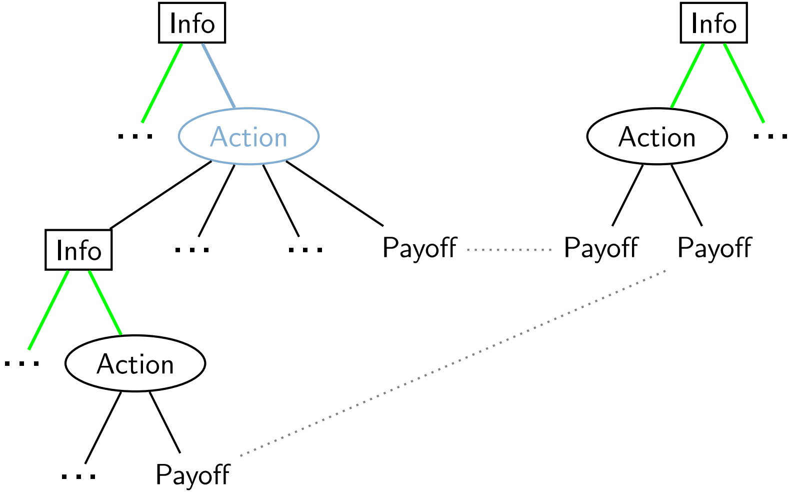
Ways to improve convergence

- Scaling regret
- Counterfactual best response on subtrees with zero reach probability.
- Pruning

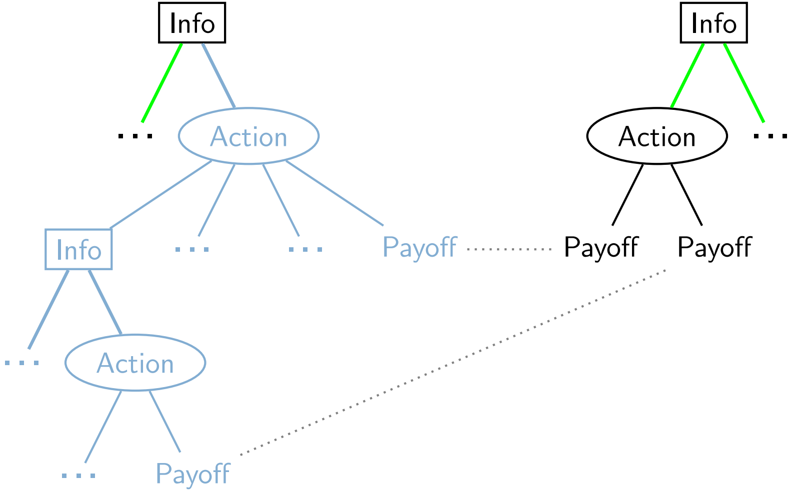
Pruning

- Parts of the gametree which are not reached by strategies do not need to be traversed.

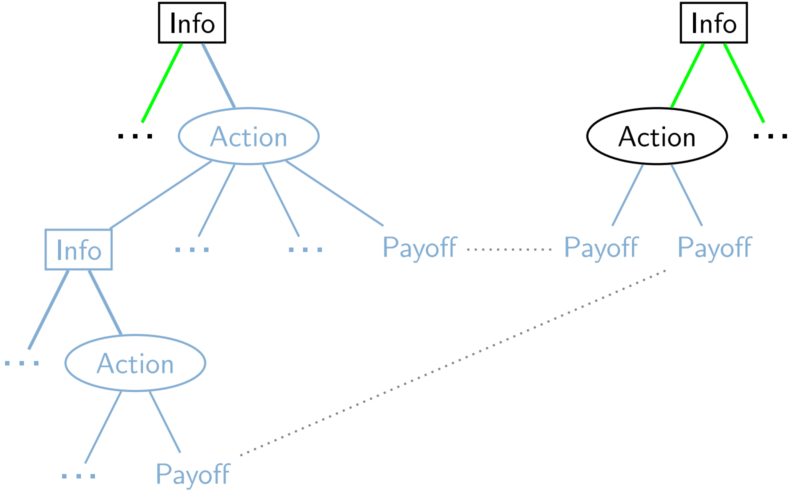
Pruning


Info
Action
Info
Payoff
Payoff
Payoff
Info
Action
Action
Payoff

Pruning


Info
...
Action
Info
...
...
Payoff
...
Action
...
Payoff
Info
Action
...
Payoff
Payoff

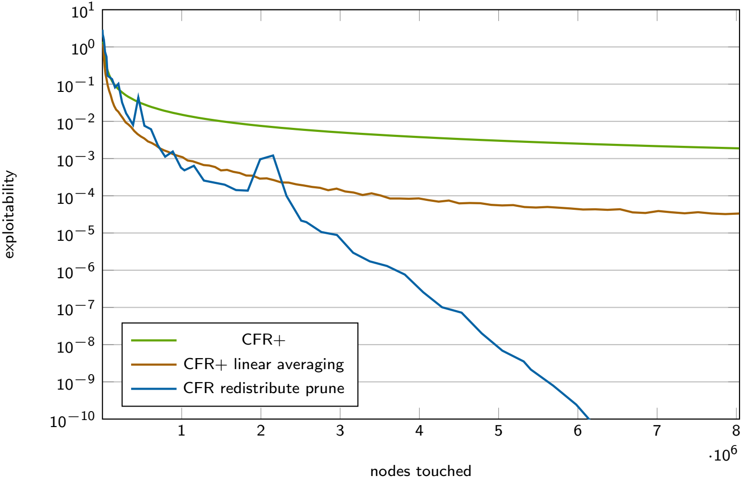
Pruning


Info
Info
...
Action
Action
...
Info
...
...
Payoff
Payoff
Payoff
...
Action
...
Payoff

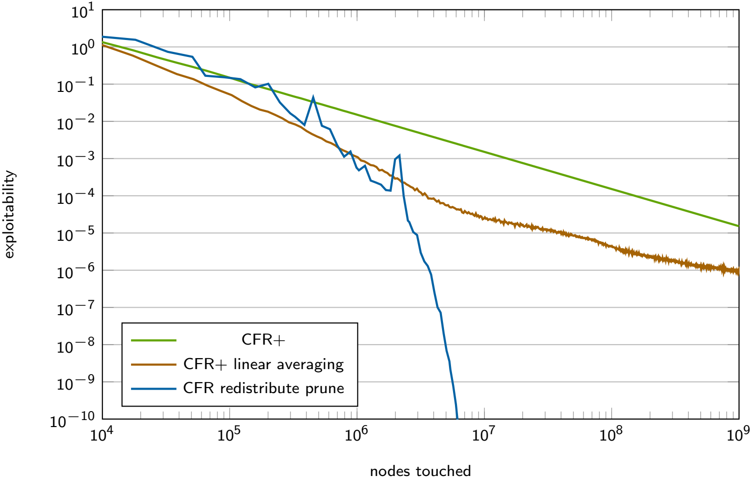
Pruning

- Parts of the game tree which are not reached by strategies do not need to be traversed.
- Prune over full block of iterations
- Condition for pruning given by average strategy $\bar{\sigma}$ of previous block. Action is pruned if both of the following hold:
  - $\bar{\sigma}(a) = 0$
  - $r_{\bar{\sigma}}(a) < 0$
- After the block, $\bar{\sigma}$ is set to counterfatual best response on the pruned parts of the game tree.
- Pruning too aggressive, have to include some short unpruned blocks of iterations.

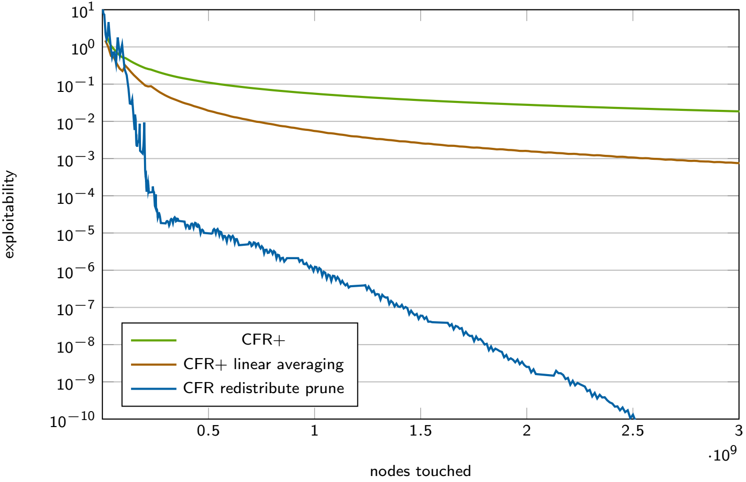
Leduc Hold'em



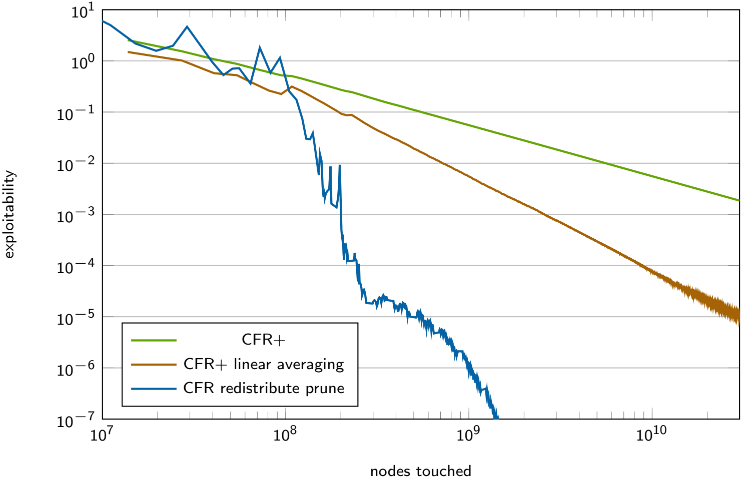
Leduc Hold'em



Leduc 5


exploitability
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
0.5
1
1.5
2
2.5
3
$\cdot 10^{9}$
nodes touched
CFR+
CFR+ linear averaging
CFR redistribute prune

Leduc 5


exploitability
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{7}
10^{8}
10^{9}
10^{10}
nodes touched
CFR+
CFR+ linear averaging
CFR redistribute prune

Michael Schubert
M. Sc. Mathematics
Software Consultant

TNG TECHNOLOGY CONSULTING

TNG Technology Consulting GmbH
Betastr. 13a
85774 Unterföhring

Tel.
Fax
Mobil +49 174 3417070

michael.schubert@tngtech.com

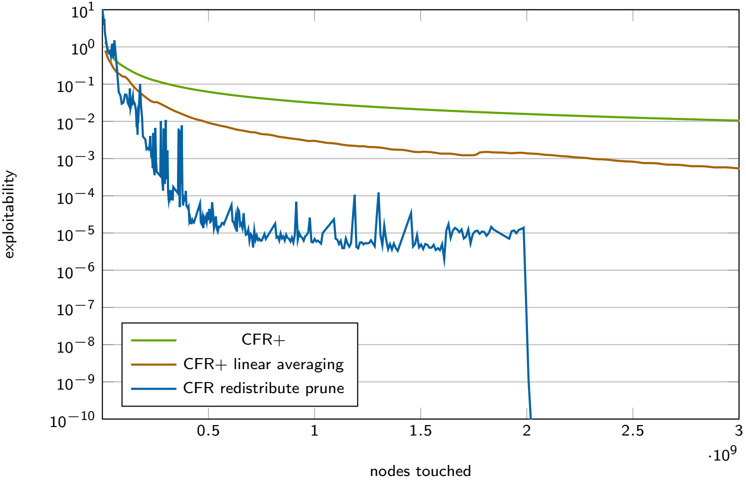
Readings:

On CFR: M. Zinkevich et. al. 2007: Regret Minimization in Games with Incomplete Infromation http://martin.zinkevich.org/publications/regretpoker.pdf

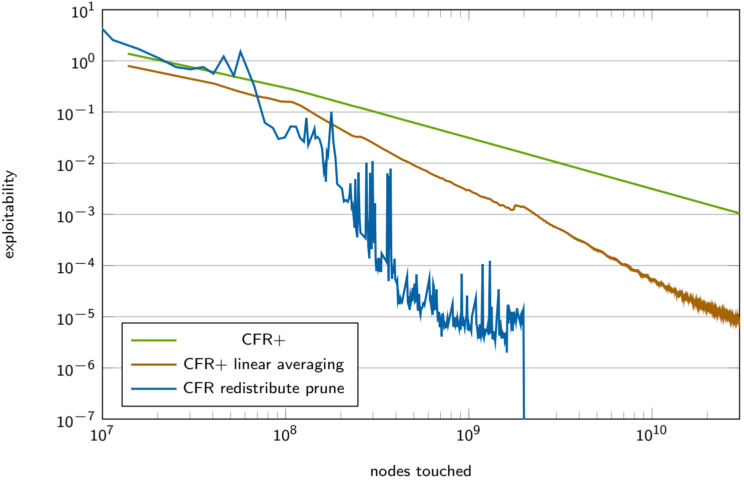
On CFR+: O. Tammelin 2014: Solving Large Imperfect Information Games Using CFR+ https://arxiv.org/abs/1407.5042

On pruning: N. Brown, T. Sandholm 2015: Regret-Based Pruning in Extensive-Form Games https://www.cs.cmu.edu/∼noamb/papers/15-NIPS-Regret-Based.pdf

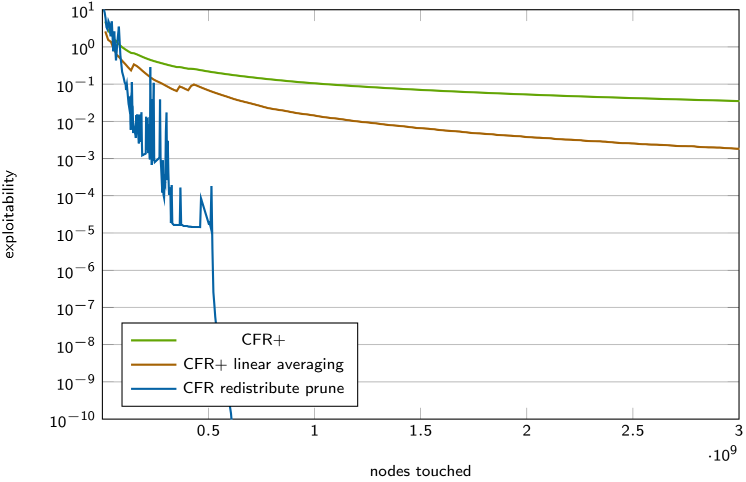
More Plots - Leduc 5 (0.5, 1, 2, 4, 8)



More Plots - Leduc 5 (0.5, 1, 2, 4, 8)



More Plots - Leduc 5 (2, 4, 8, 16, 32)



More Plots - Leduc 5 (2, 4, 8, 16, 32)