Efficient Monte Carlo Counterfactual Regret Minimization in Games with Many Player Actions

Richard Gibson, Neil Burch, Marc Lanctot, and Duane Szafron

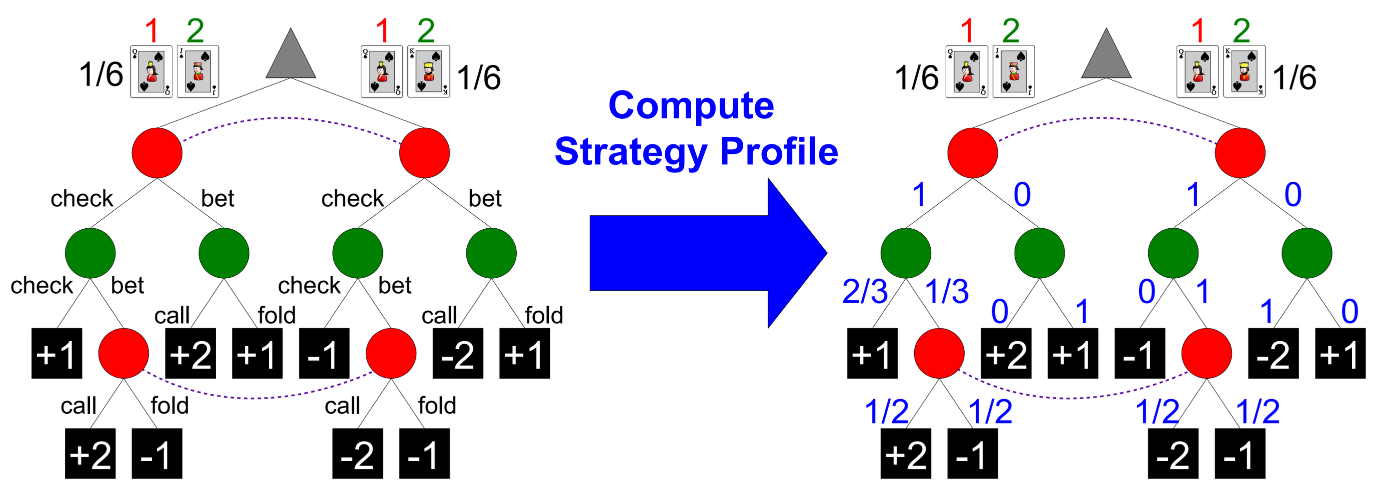
Computing Science Department, University of Alberta, Canada

Poster available on-line at http://cs.ualberta.ca/~rggibson/

1. MOTIVATION

Goal: Find solutions to large 2-player zero-sum imperfect information games.

Example: Kuhn Poker (player 1 dealt Queen)



We seek a **Nash equilibrium profile** (or as close to Nash as possible)

Applications: Airport security, insulin scheduling for diabetes patients, **beat humans at Texas Hold'em poker**.

NOTATION AND DEFINITIONS

$\sigma = (\sigma_1, \sigma_2)$: **strategy profile**, a function mapping each information set to a probability distribution over actions

$u_i(\sigma)$: **expected utility** for player i, assuming players play according to σ

exploitability$(\sigma) = \dfrac{\max_{\sigma_2'} u_2(\sigma_1, \sigma_2') + \max_{\sigma_1'} u_1(\sigma_1', \sigma_2)}{2}$: maximum amount σ loses to a worst-case opponent

A strategy profile σ is an ϵ**-Nash equilibrium** if exploitability$(\sigma) \le \epsilon$

T: number of iterations

$R_1^T = \max_{\sigma_1'} \sum_{t=1}^{T} u_1(\sigma_1', \sigma_2^t) - u_1(\sigma_1^t, \sigma_2^t)$: **regret** for player 1 after T iterations

$\mathcal{I}_i$: set of information sets for player i

$R_1^T(I) = \max_a \sum_{t=1}^{T} \pi_{-i}^{\sigma^t}(I)\left(u_1(\sigma^t_{1(I\to a)}, \sigma_2^t \mid I) - u_1(\sigma_1^t, \sigma_2^t \mid I)\right)$:

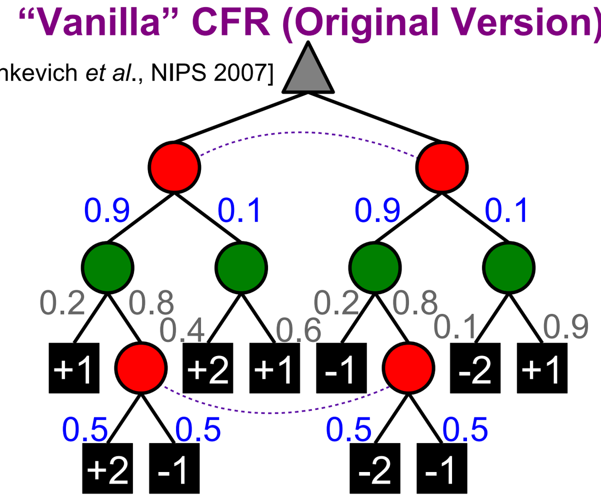
counterfactual regret for player 1 at information set I

2. BACKGROUND

Counterfactual Regret Minimization (CFR) is a state-of-the-art iterative algorithm for computing an approximate Nash equilibrium.

"Vanilla" CFR (Original Version)

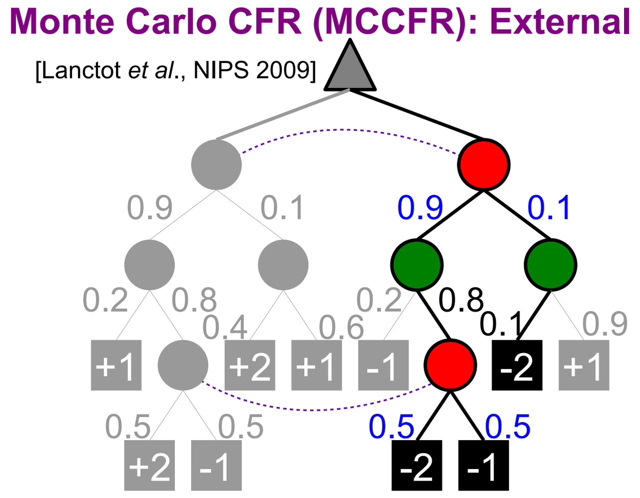
[Zinkevich *et al.*, NIPS 2007]



Traverse entire tree each iteration.
- slow iterations
- few iterations required

Monte Carlo CFR (MCCFR): External Sampling

[Lanctot *et al.*, NIPS 2009]



Only traverse a sampled subtree.
- fast iterations
- many iterations required

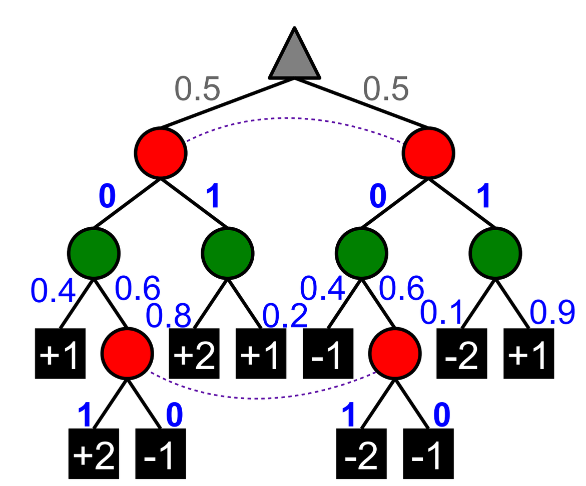
Output: $\bar{\sigma}^T = \dfrac{\sigma^1 + \sigma^2 + \cdots + \sigma^T}{T}$, the **average strategy profile.**

Regret Bound: $R_i^T \le \sum_{I \in \mathcal{I}_i} R_i^{T,+}(I) \le C\sqrt{T}$ [Zinkevich *et al.*, NIPS 2007]

Fact: $\dfrac{R_1^T}{T}, \dfrac{R_2^T}{T} < \dfrac{\epsilon}{2} \Rightarrow \bar{\sigma}^T$ is an ϵ-Nash equilibrium.

3. NEW THEORETICAL RESULT

Let σ_i^* be a best response to $\bar{\sigma}_{-i}^T$:



New Regret Bound: $R_i^T = \sum_{I \in \mathcal{I}_i} \pi_i^{\sigma^*}(I) R_i^T(I) \le C^* \sqrt{T}, C^* \le C,$

where $\pi_i^{\sigma^*}(I)$ is the probability σ_i^* plays to reach I.

Observation 1: Regret only depends on counterfactual regret $R_i^T(I)$ at information sets I that σ_i^* plays to reach.

Observation 2: $\bar{\sigma}_i^T \to \sigma_i^*$ as $T \to \infty$

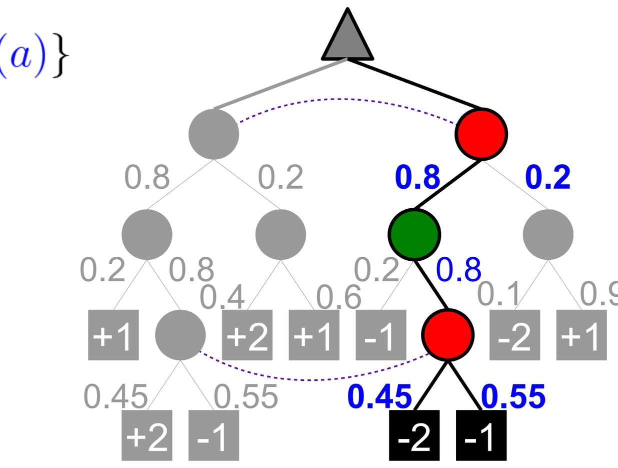
4. NEW SAMPLING ALGORITHM

Main Contribution: New MCCFR sampling algorithm, **Average Strategy Sampling**, that samples a subset of the current player's actions according to the player's average strategy.

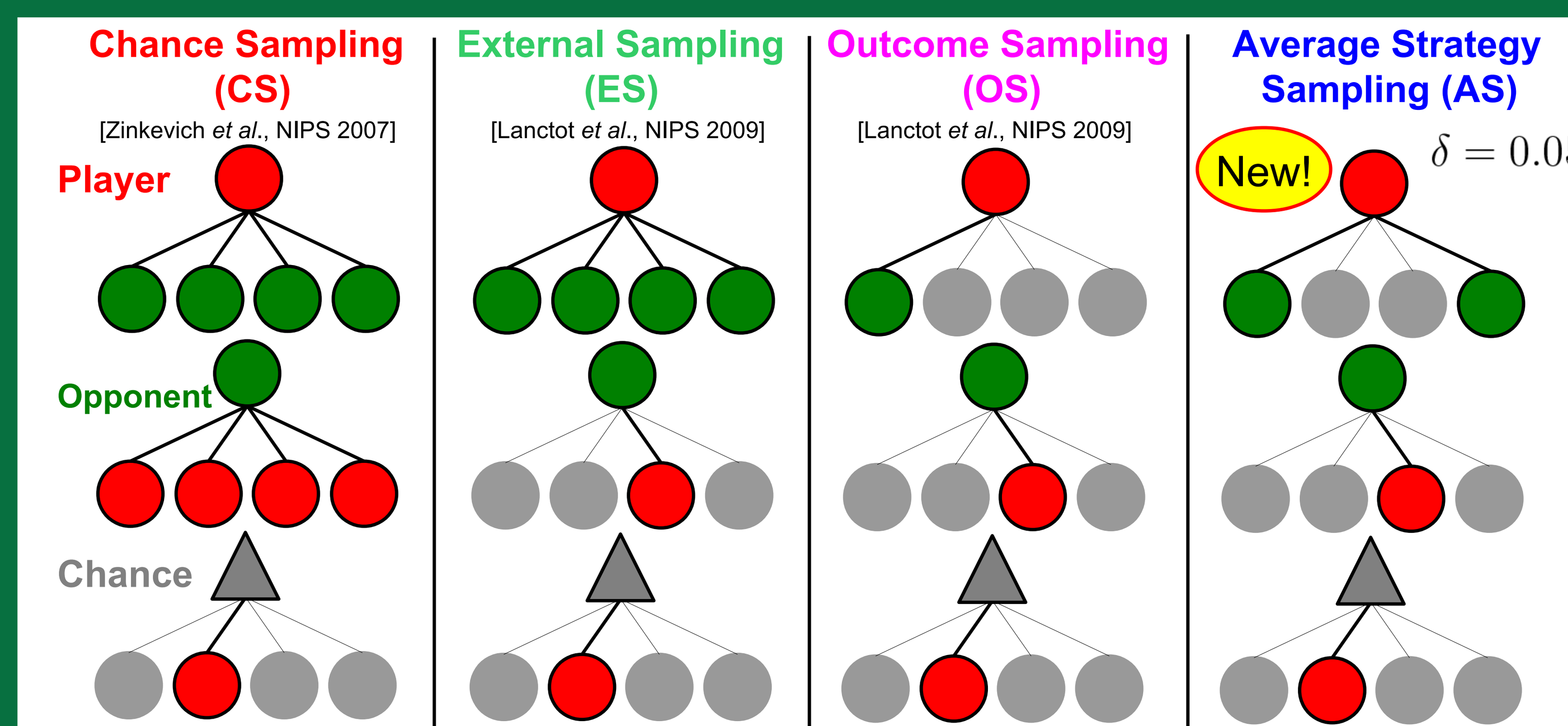
$\mathrm{Prob}[\text{sample action } a] \approx \max\{\delta, \bar{\sigma}_i^T(a)\}$

exploration parameter δ

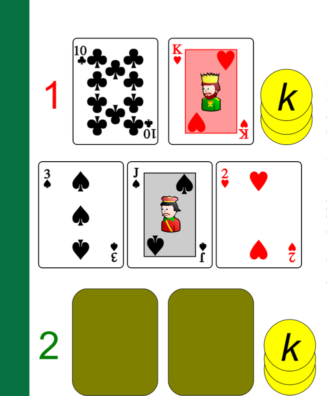
- even faster iterations
- focus effort more on where we will play in practice

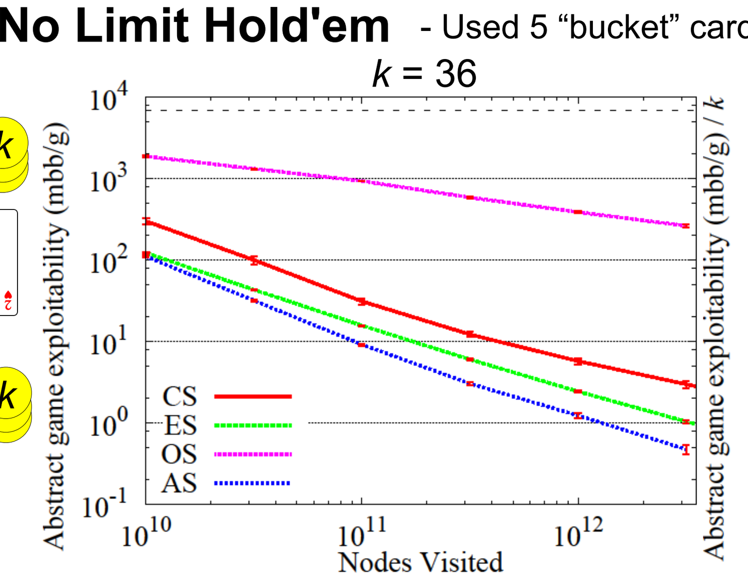


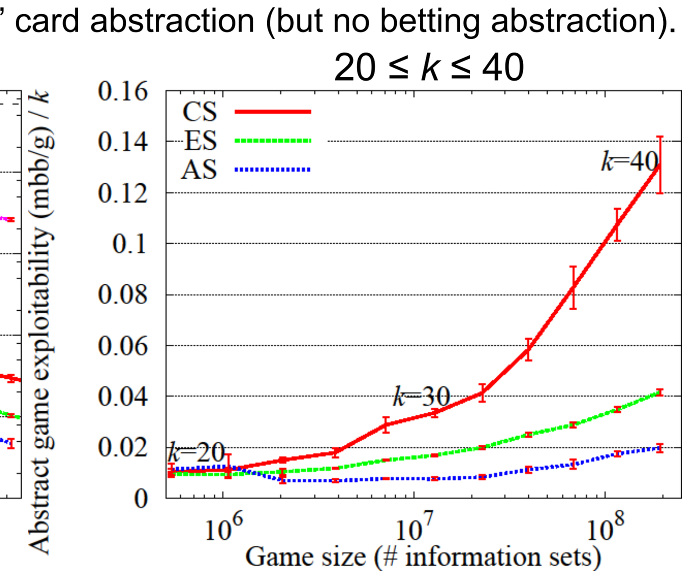
5. EXPERIMENTAL RESULTS



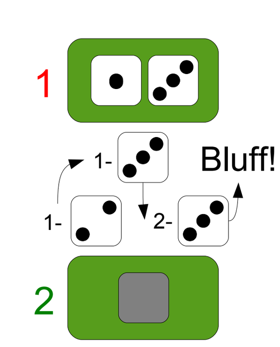
2-Round No Limit Hold'em - Used 5 "bucket" card abstraction (but no betting abstraction).

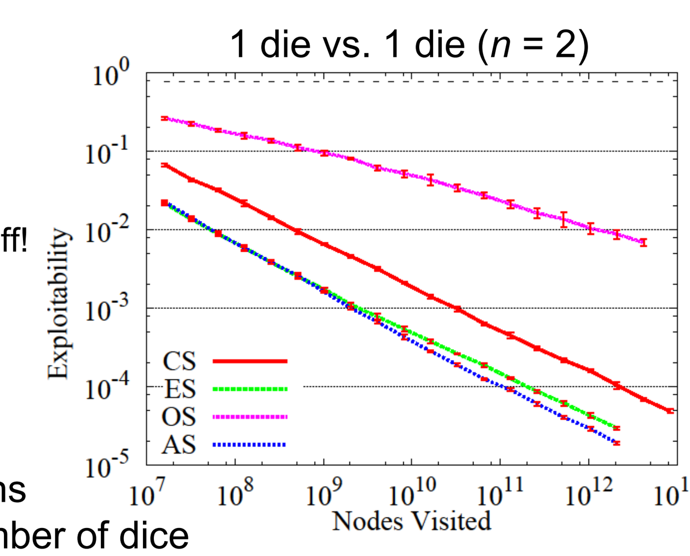


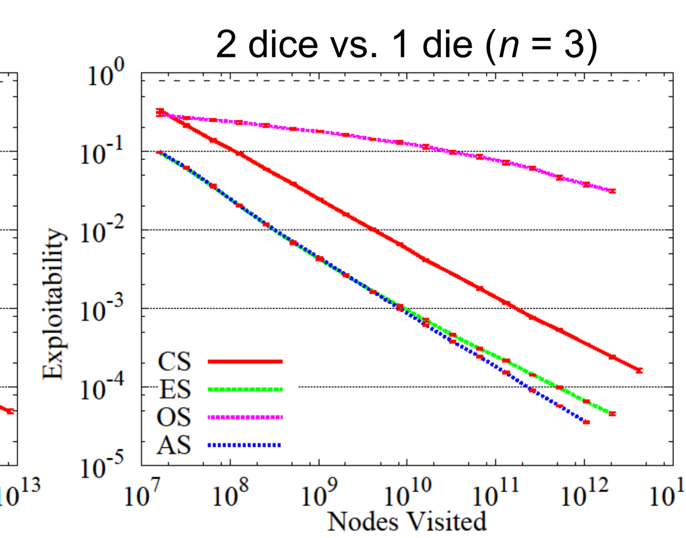




Bluff







RESEARCH SUPPORTED BY:

compute • calcul CANADA, WestGrid, Alberta Innovates, NSERC CRSNG, UNIVERSITY OF ALBERTA