Generalized Sampling and Variance in Counterfactual Regret Minimization

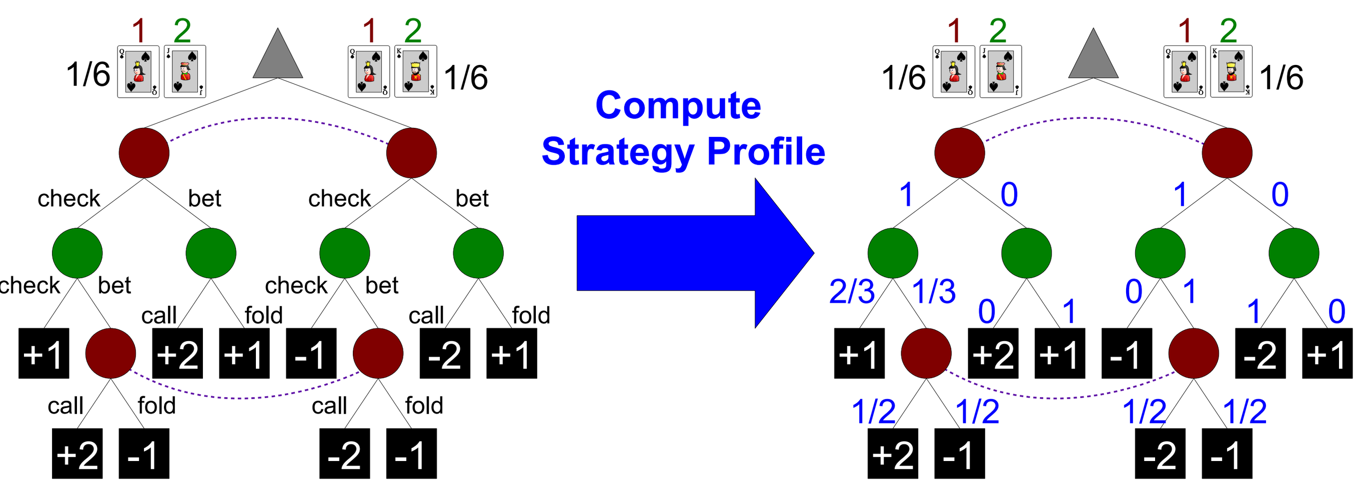
Richard Gibson, Marc Lanctot, Neil Burch, Duane Szafron, and Michael Bowling

Computing Science Department, University of Alberta, Canada

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: Security games, sports strategy, **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$(\sigma) = \frac{max_{\sigma_2'} u_2(\sigma_1, \sigma_2') + max_{\sigma_1'} u_1(\sigma_1', \sigma_2)}{2}$:

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

$|I_i|$: number of information sets for player i

$|A_i|$: maximum number of actions available at an information set for player i

$\Delta_i, \tilde{\Delta}_i, \hat{\Delta}_i$: largest possible difference between two v calculations for player i

RESEARCH SUPPORTED BY:

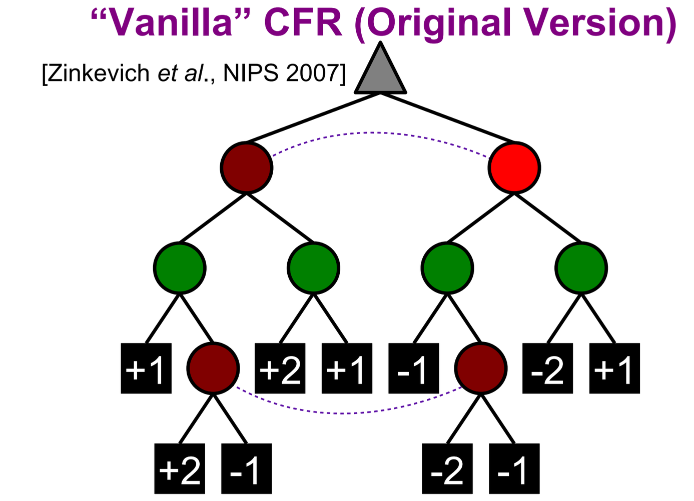
Compute • Calcul Canada, WestGrid, Alberta Innovates, NSERC CRSNG, University of Alberta

2. BACKGROUND

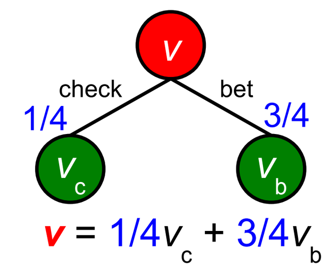
Counterfactual Regret Minimization (CFR) is a state-of-the-art, iterative algorithm for computing ϵ-Nash equilibria in large imperfect information games.

"Vanilla" CFR (Original Version)

[Zinkevich *et al.*, NIPS 2007]



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



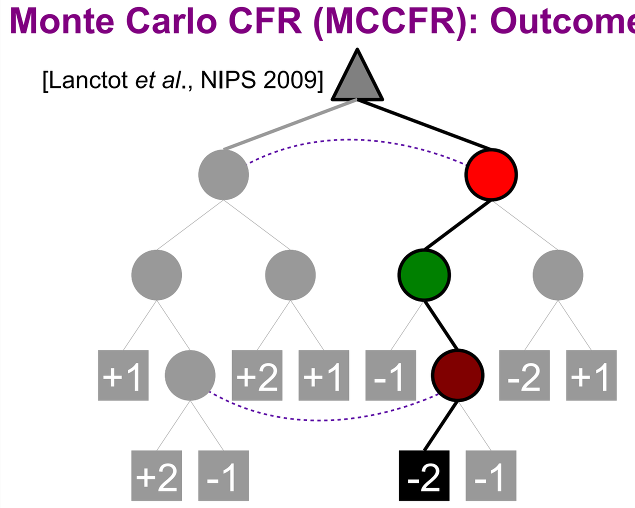
$v = 1/4 v_c + 3/4 v_b$

- v is the true expected value at this node.
- Action probabilities updated based on v, v_c, and v_b.

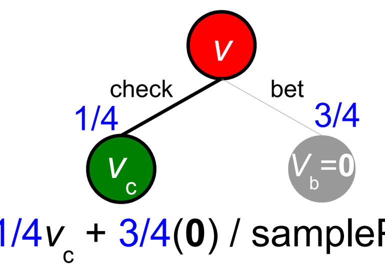
$$\frac{R_i^T}{T} \le \frac{\Delta_i |I_i| \sqrt{|A_i|}}{\sqrt{T}}$$

Monte Carlo CFR (MCCFR): Outcome Sampling

[Lanctot *et al.*, NIPS 2009]



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



$v = 1/4 v_c + 3/4(\mathbf{0})$ / sampleProb(check)

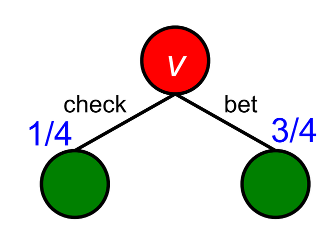
- v is an unbiased estimate of the true expected value.
- Variance introduced through sampling.

$$\frac{R_i^T}{T} \le \left(\tilde{\Delta}_i + \frac{\sqrt{2}\,\tilde{\Delta}_i}{\sqrt{p}}\right) \frac{|I_i| \sqrt{|A_i|}}{\sqrt{T}}$$

It is well-known that if $\frac{R_i^T}{T} < \frac{\epsilon}{2}$ for i = 1,2, then the average of the strategy profiles generated is an ϵ-Nash equilibrium.

3. NEW THEORETICAL RESULTS

Contribution 1: We generalize **MCCFR** by showing that v can be ANY estimate of the true expected value at a given node:



v = any estimate of the true expected value at this node

- strategies updated based on v as before.
- convergence to equilibrium achieved when v is unbiased.

Contribution 2: We provide a bound on the average regret in terms of the variance, covariance, and bias of v. When v is unbiased, we have the following bound on the convergence rate:

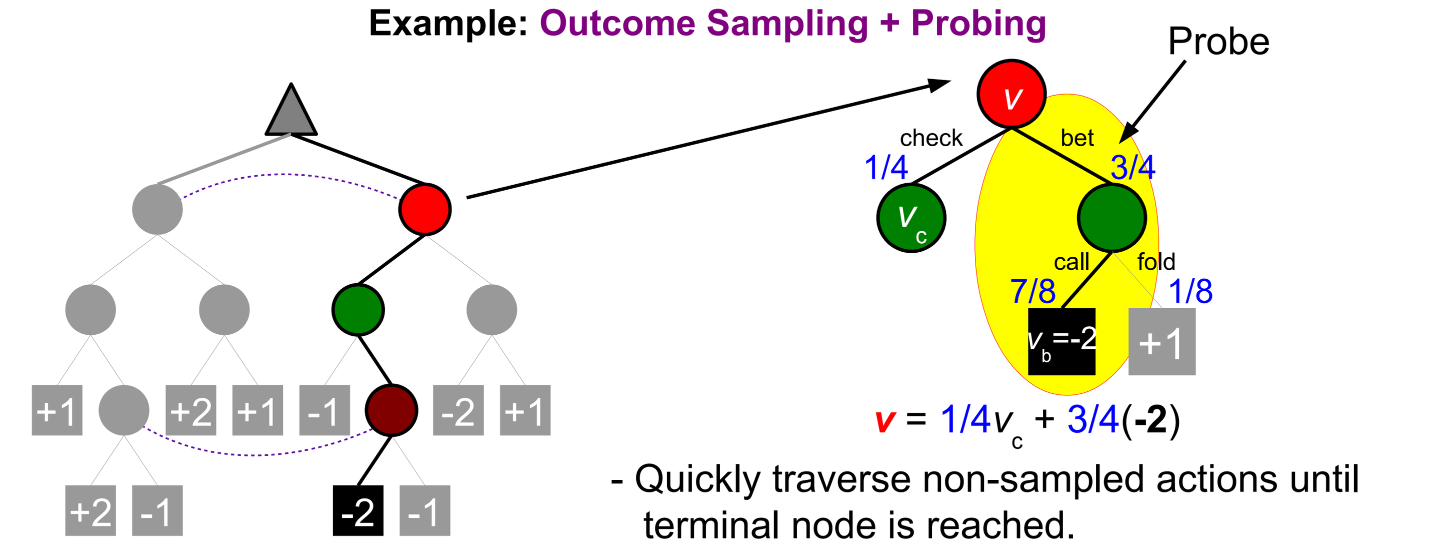
Theorem: For $p \in (0,1]$, if v is unbiased, then with probability at least $1-p$,

$$\frac{R_i^T}{T} \le \left(\hat{\Delta}_i + \frac{\sqrt{\textbf{Var}[\,v\,]}}{\sqrt{p}}\right) \frac{|I_i| \sqrt{|A_i|}}{\sqrt{T}}$$

4. NEW SAMPLING ALGORITHM

Contribution 3: We introduce a new CFR sampling variant called **Probing** that provides lower variance estimates v when combined with an **MCCFR** sampling scheme.

Example: Outcome Sampling + Probing

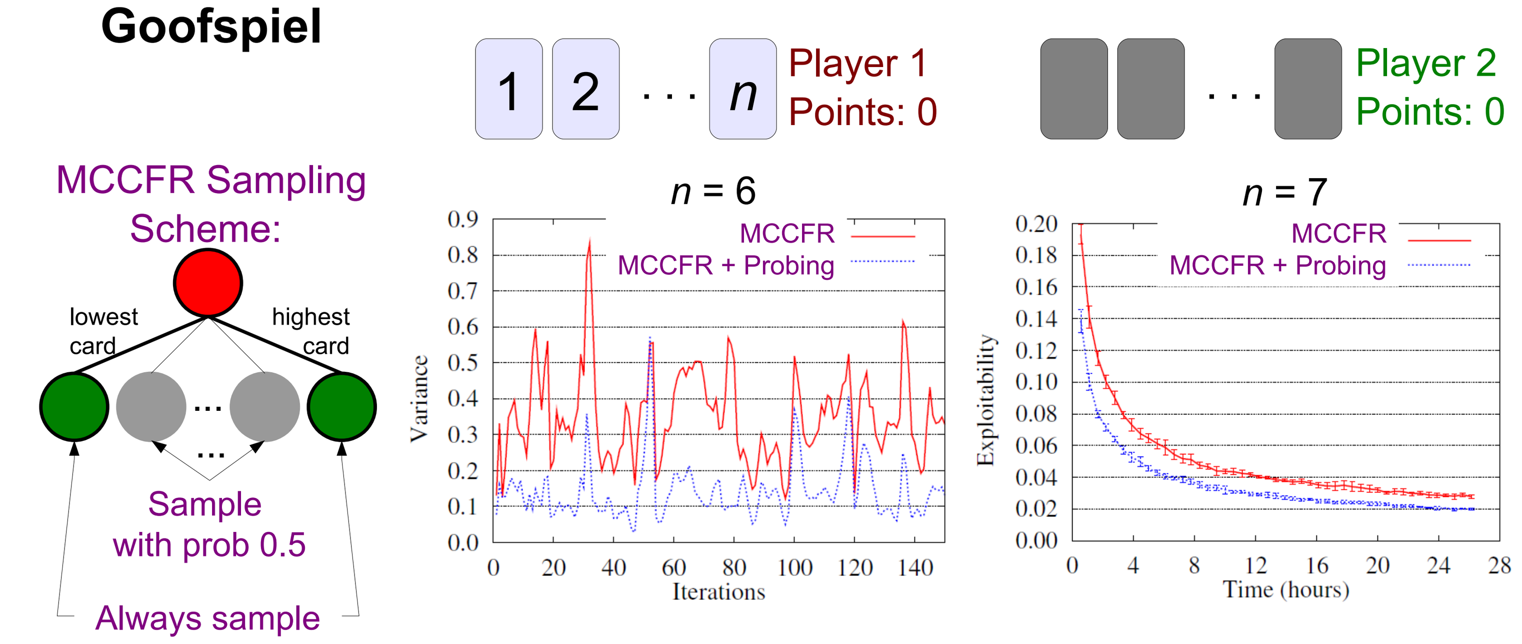


$v = 1/4 v_c + 3/4(\mathbf{-2})$

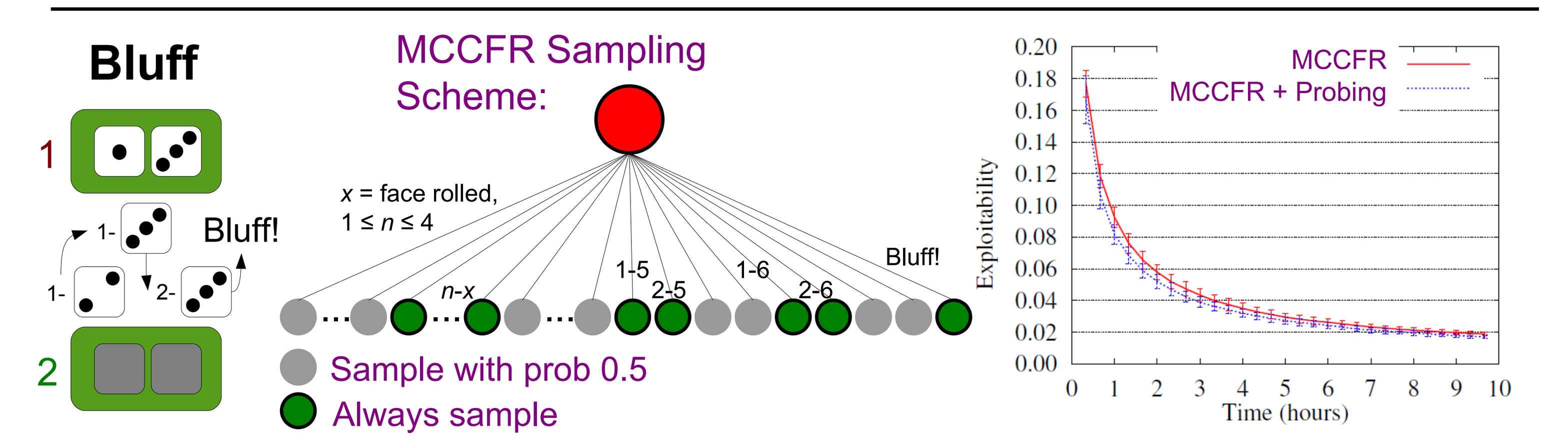
- Quickly traverse non-sampled actions until terminal node is reached.
- v remains unbiased with reduced variance.
- New theory suggests fewer iterations required, but at the cost of slightly slower iterations.

5. EXPERIMENTAL RESULTS

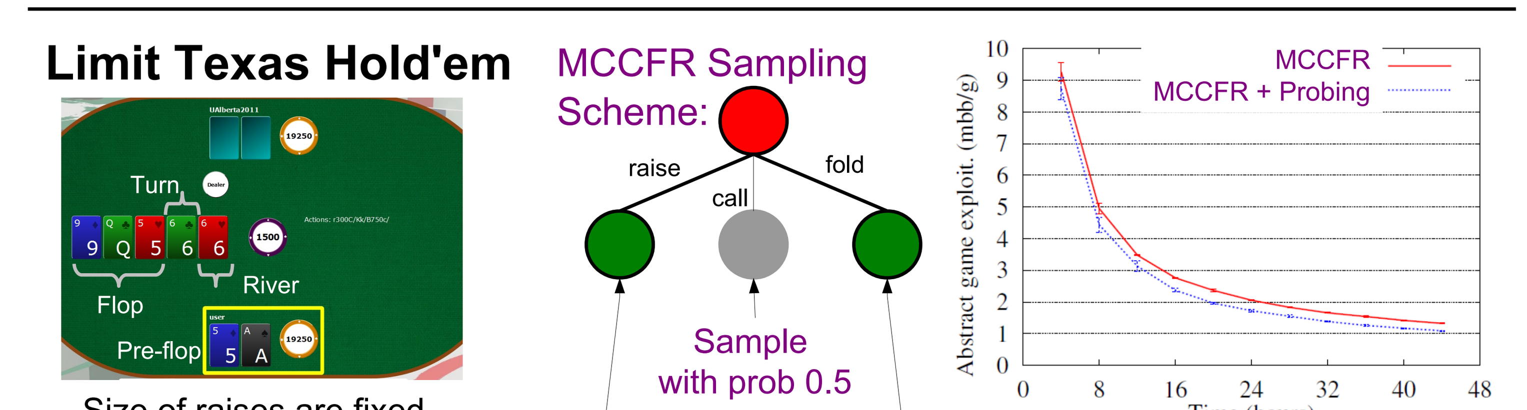
Goofspiel



Bluff



Limit Texas Hold'em



- Size of raises are fixed.
- Used 10 "bucket" card abstraction.