Rational Monte Carlo Tree Search

David Tolpin, Solomon Eyal Shimony

Ben-Gurion University of the Negev
Beer Sheva, Israel

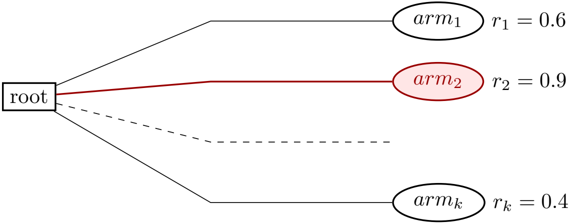
June 1, 2011

UCT

UCT (Upper Confidence Bounds on Trees) is popular for Monte Carlo search in large trees:

- Selects an action a_i that maximizes $\overline{r}_i + C\sqrt{\frac{\log n}{n_i}}$
- Chooses a non-optimal action $n_i \geq \rho \log n$ times (ρ is some constant).
- Based on UCB1, that achieves logarithmic regret for multi-armed bandits.
- But **no bandits in Monte Carlo: no reward** is given for **sampling** a good action.

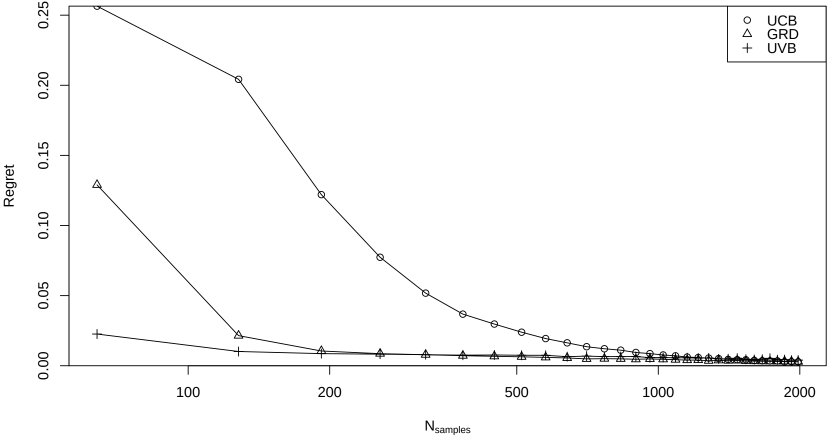
Doing better than UCT on sets



When an arm is selected based on the **sample mean**:

- Regret of UCB decreases *polynomially* with n.
- Regret of ϵ-greedy decreases *exponentially* with n.
- Regret of UVB: $\max V_i, \ V_{i_{best}} = \frac{1-1/k}{n_{i_{best}}}, \quad V_{i_{other}} = \frac{1/k}{n_{i_{other}}}$ decreases exponentially with n, faster than ϵ-greedy.

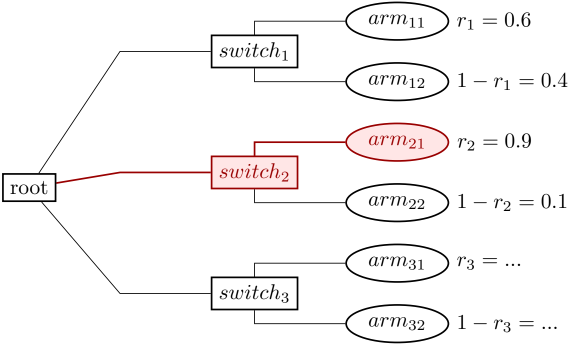
UCB vs. ϵ-greedy vs UVB



64 Bernoulli arms, randomly generated

Doing Better Than UCT on Trees

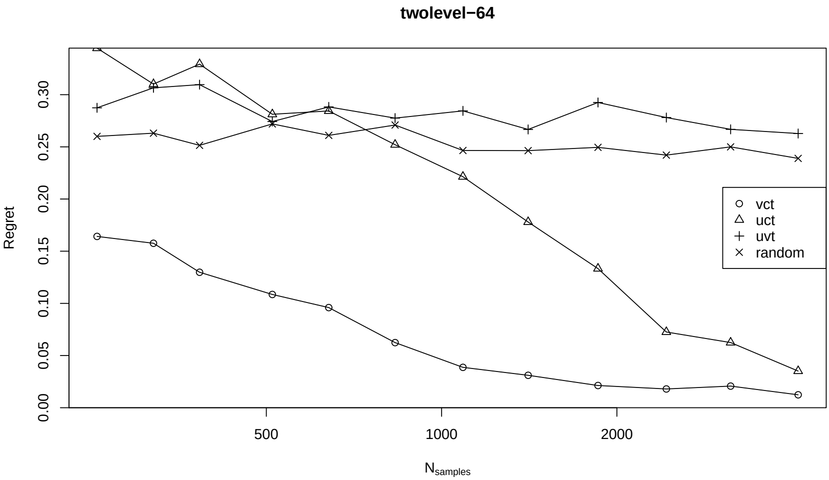
Uniform sampling is useless in this tree:



Rational sampling:

- first, choose an action that maximizes VOI (UVB);
- then, choose actions that maximize average reward (UCB).

UVT vs. VCT (UVB+UCT) vs. UCT



64 Bernoulli arms, randomly generated