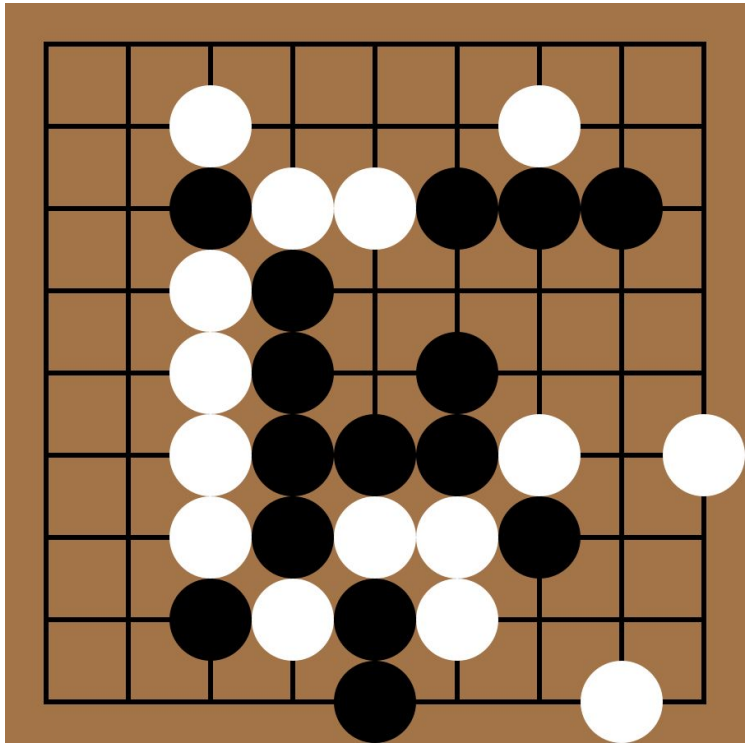
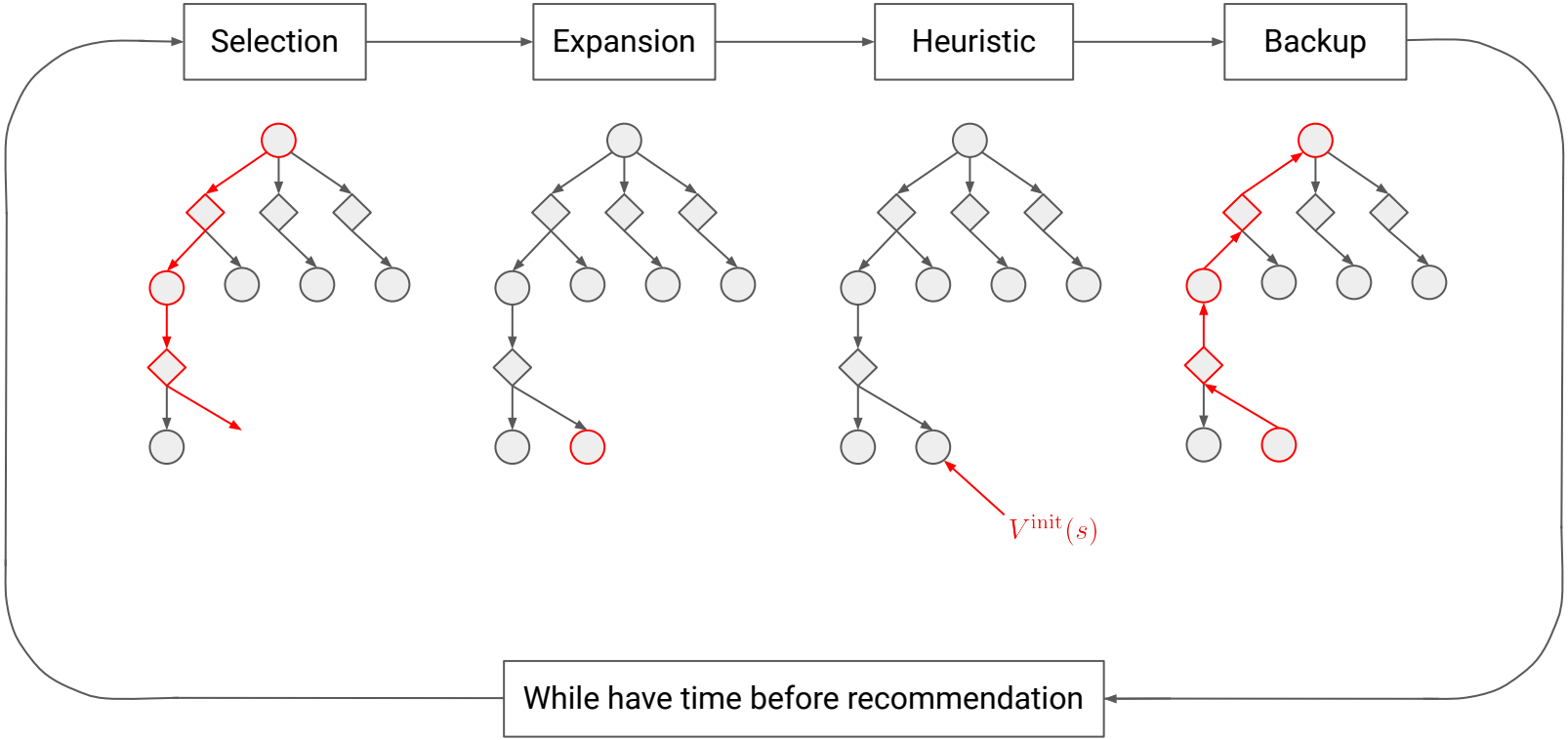
Monte-Carlo Tree Search with Boltzmann Exploration

Michael Painter, Mohamed Baioumy, Bruno Lacerda, Nick Hawes

Monte-Carlo Tree Search



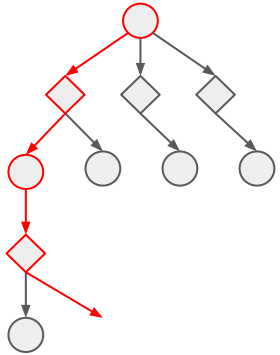
Monte-Carlo Tree Search



UCB Applied to Trees (UCT)

GOAL-ORIENTED AUTONOMOUS LONG-LIVED SYSTEMS
OXFORD ROBOTICS INSTITUTE

Selection



$Q(s, a)$
- Q-Value Estimate

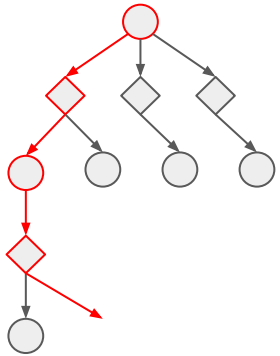
$N(s), N(s, a)$
- Visit Counts

UCB Applied to Trees (UCT)

Action selection:

$$\pi_{\mathrm{UCT}}(s) = \mathrm{argmax}_a \left[Q(s,a) + c\frac{\log(N(s))}{N(s,a)} \right]$$

Selection



$Q(s,a)$

- Q-Value Estimate

$N(s), N(s,a)$

- Visit Counts

UCB Applied to Trees (UCT)

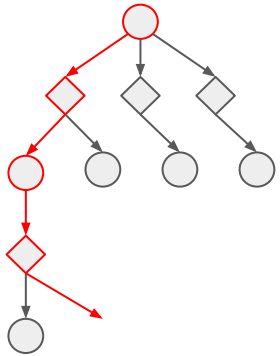
Action selection:

$$\pi_{\mathrm{UCT}}(s) = \mathrm{argmax}_a \left[Q(s,a) + c\frac{\log(N(s))}{N(s,a)} \right]$$

Recommendation policy:

$$\psi_{\mathrm{UCT}}(s) = \mathrm{argmax}_a Q(s,a)$$

Selection



$Q(s,a)$

- Q-Value Estimate

$N(s), N(s,a)$

- Visit Counts

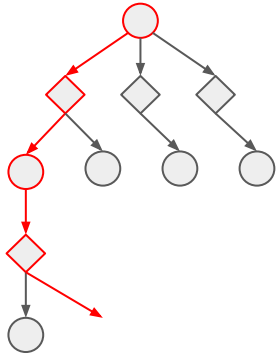
Boltzmann Tree Search (BTS)

GOAL-ORIENTED AUTONOMOUS LONG-LIVED SYSTEMS
OXFORD ROBOTICS INSTITUTE

Action Selection:

$$\pi_{\mathrm{BTS}}(a|s) \propto \exp\left(\frac{Q(s,a)}{\alpha}\right)$$

Selection



$Q(s,a)$

- Q-Value Estimate

$N(s), N(s,a)$

- Visit Counts

Boltzmann Tree Search (BTS)

GOAL-ORIENTED AUTONOMOUS LONG-LIVED SYSTEMS
OXFORD ROBOTICS INSTITUTE

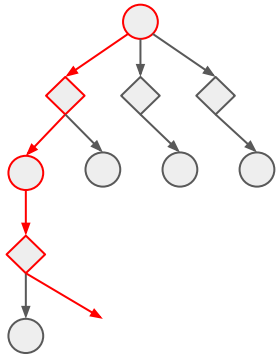
Action Selection:

$$\pi_{\mathrm{BTS}}(a|s) \propto \exp\left(\frac{Q(s,a)}{\alpha}\right)$$

Recommendation Policy:

$$\psi_{\mathrm{BTS}}(s) = \mathrm{argmax}_a Q(s,a)$$

Selection



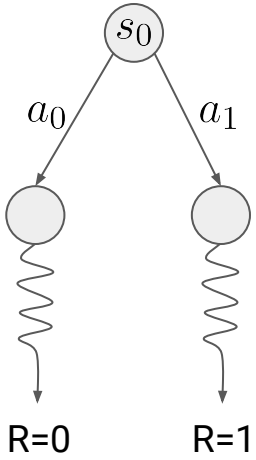
$Q(s,a)$

- Q-Value Estimate

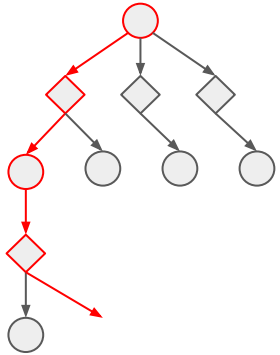
$N(s), N(s,a)$

- Visit Counts

Exploration in Action Selection



Selection



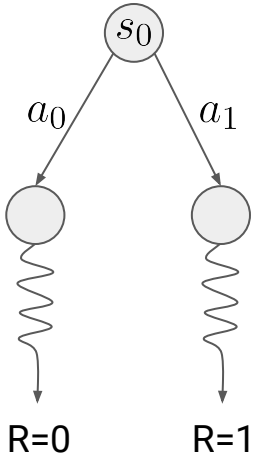
$Q(s, a)$

- Q-Value Estimate

$N(s), N(s, a)$

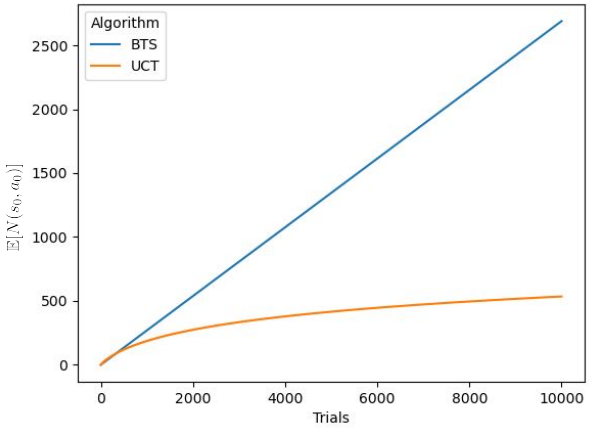
- Visit Counts

Exploration in Action Selection

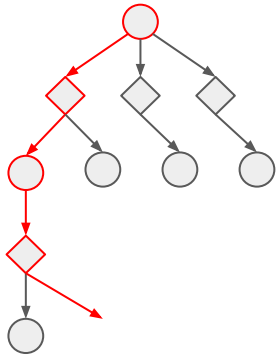


$$\pi_{\mathrm{BTS}}(a|s) \propto \exp\left(\frac{Q(s,a)}{\alpha}\right)$$

$$\pi_{\mathrm{UCT}}(s) = \mathrm{argmax}_a \left[Q(s,a) + c\frac{\log(N(s))}{N(s,a)}\right]$$



Selection



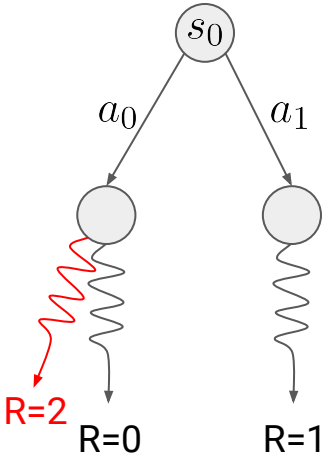
$Q(s,a)$

- Q-Value Estimate

$N(s), N(s,a)$

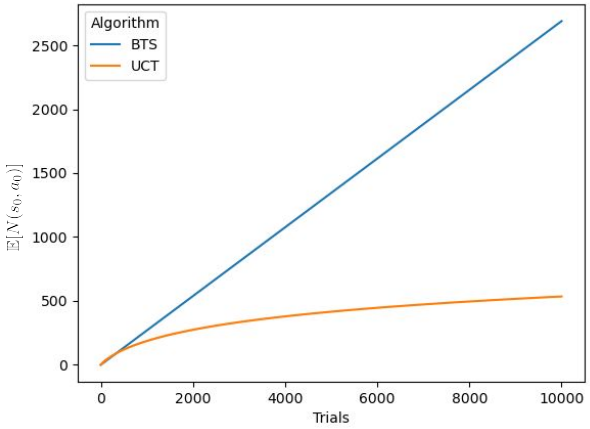
- Visit Counts

Exploration in Action Selection

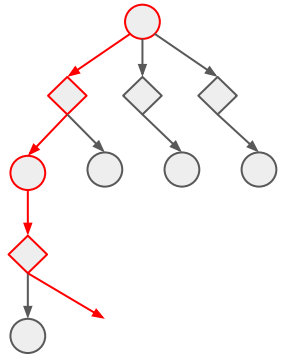


$$\pi_{\mathrm{BTS}}(a|s) \propto \exp\left(\frac{Q(s,a)}{\alpha}\right)$$

$$\pi_{\mathrm{UCT}}(s) = \mathrm{argmax}_a \left[Q(s,a) + c\frac{\log(N(s))}{N(s,a)}\right]$$



Selection



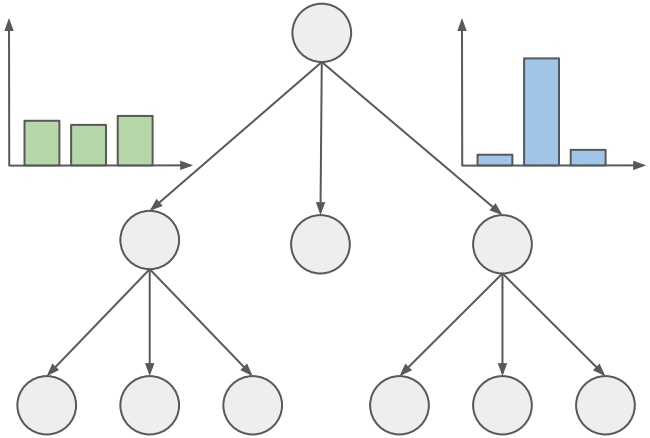
$Q(s,a)$

- Q-Value Estimate

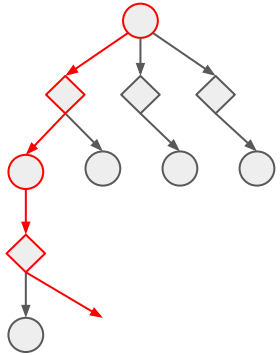
$N(s), N(s,a)$

- Visit Counts

Entropy in Tree Search



Selection



$Q(s, a)$

- Q-Value Estimate

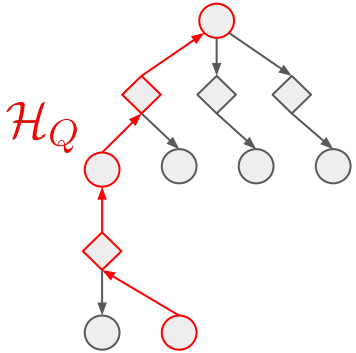
$N(s), N(s, a)$

- Visit Counts

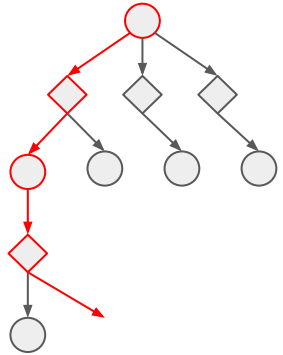
Decaying ENtropy Tree Search (DENTS)

GOAL-ORIENTED AUTONOMOUS LONG-LIVED SYSTEMS
OXFORD ROBOTICS INSTITUTE

Backup



Selection



$Q(s, a)$

- Q-Value Estimate

$N(s), N(s, a)$

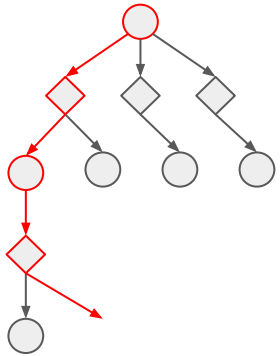
- Visit Counts

Decaying ENtropy Tree Search (DENTS)

Action Selection:

$$\pi_{\text{DENTS}}(a|s) \propto \exp\left(\frac{Q(s,a)+\beta(N(s))\mathcal{H}_Q(s,a)}{\alpha}\right)$$

Selection



$Q(s,a)$

- Q-Value Estimate

$N(s), N(s,a)$

- Visit Counts

Decaying ENtropy Tree Search (DENTS)

GOAL-ORIENTED AUTONOMOUS LONG-LIVED SYSTEMS
OXFORD ROBOTICS INSTITUTE

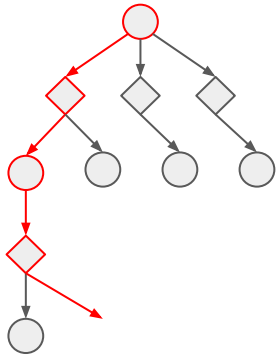
Action Selection:

$$\pi_{\text{DENTS}}(a|s) \propto \exp\left(\frac{Q(s,a)+\beta(N(s))\mathcal{H}_Q(s,a)}{\alpha}\right)$$

Recommendation Policy:

$$\psi_{\text{DENTS}}(s) = \text{argmax}_a Q(s,a)$$

Selection



$Q(s,a)$
- Q-Value Estimate

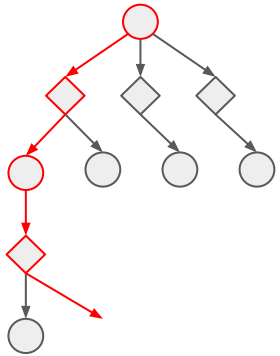
$N(s), N(s,a)$
- Visit Counts

Alias Method

Distribution with A actions:

- Build alias table with O(A) cost
- Sample actions in O(1) time

Selection



$Q(s, a)$

- Q-Value Estimate

$N(s), N(s, a)$

- Visit Counts

Alias Method

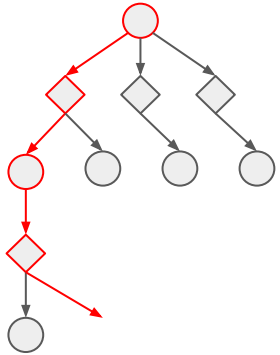
Distribution with A actions:

- Build alias table with O(A) cost
- Sample actions in O(1) time

Faster action selection using stochastic policies:

1. If N(s) is a multiple of A:
   Build new alias table
2. Sample action using alias table

Selection



$Q(s, a)$

- Q-Value Estimate

$N(s), N(s, a)$

- Visit Counts

Final Comments

Discussed:

- How Boltzmann policies explore more during action selection
- Using entropy as an exploration bonus to motivate DENTS
- Faster action sampling using the Alias method

Final Comments

Discussed:

- How Boltzmann policies explore more during action selection
- Using entropy as an exploration bonus to motivate DENTS
- Faster action sampling using the Alias method

See paper for:

- Convergence results for BTS and DENTS
- Discussing setting parameters in BTS, DENTS and related algorithms
- Empirically demonstrating these benefits in toy environments and Go