GROUPING NODES WITH KNOWN VALUE DIFFERENCES: A LOSSLESS UCT-BASED ABSTRACTION ALGORITHM

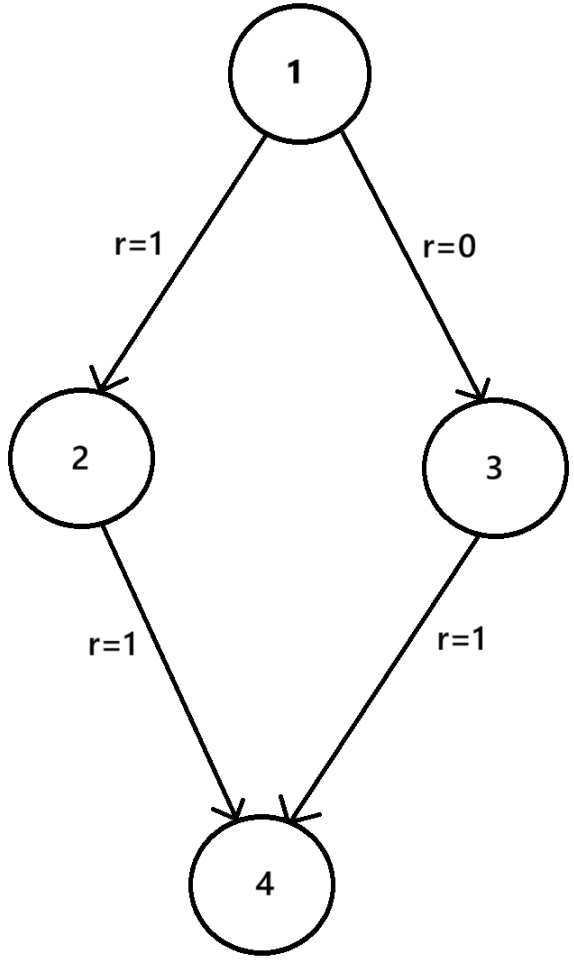
Robin Schmöcker, Alexander Dockhorn, Bodo Rosenhahn

- UCT‘s performance heavily depends on its tree policy:

$$\mathrm{UCB}(s,a) = \underbrace{\frac{V_{(s,a)}}{N_{(s,a)}}}_{\text{Q term}} + \underbrace{\lambda\sqrt{\frac{\log\left(\sum_{a' \in \mathbb{A}(s)} N_{(s,a')}\right)}{N_{(s,a)}}}}_{\text{Exploration term}}$$

- **Goal**: Detect groups of state-action pairs and use aggregate statistics instead of the single-node statistics

- **Known Value Differences Abstractions** can be built on an UCT search graph:



- Key **difference to ASAP**: Ignore immediate reward differences and keep track of the discrepancy

- Key **difference to ASAP**: Ignore immediate reward differences and keep track of the discrepancy
- UCT + KVDA = **KVDA-UCT**

- Key **difference to ASAP**: Ignore immediate reward differences and keep track of the discrepancy
- UCT + KVDA = **KVDA-UCT**
- Applicable to both deterministic and stochastic MDPs

- Key **difference to ASAP**: Ignore immediate reward differences and keep track of the discrepancy
- UCT + KVDA = **KVDA-UCT**
- Applicable to both deterministic and stochastic MDPs
- It can be shown that KVDA is **sound**