Enhancing Prost with Abstraction of State-Action Pairs

Tim Steindel

University of Basel

tim.steindel@stud.unibas.ch

November 28,2017

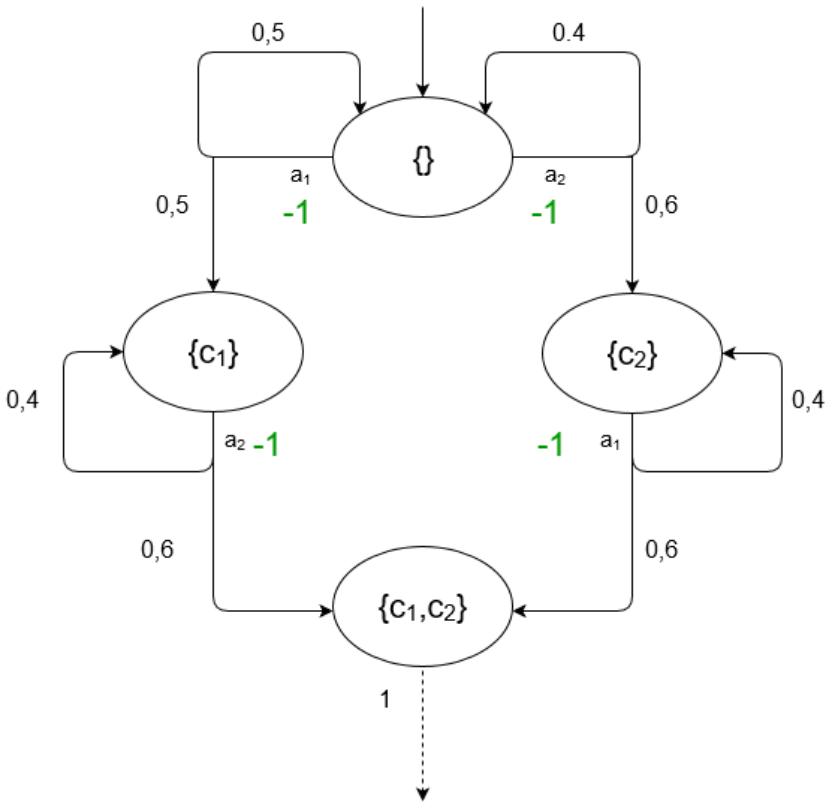
Overview

Academic advising domain

A student wants to graduate

- pass several courses
- can only take one course at a given time
- courses can overlap in content
- has a probability to pass/fail a course

Markov decision process



MDP: $\langle S, A, T, R, s_0, H \rangle$ where:

- S is a finite set of states,
- A is a finite set of actions,
- T: $S \times A \times S \to [0, 1]$ is a probability distribution,
- $R : S \times A \to \mathbb{R}$ is the reward function,
- $s_0 \in S$ is the initial state,
- $H \in \mathbb{N}$ is the finite horizon.

Search tree

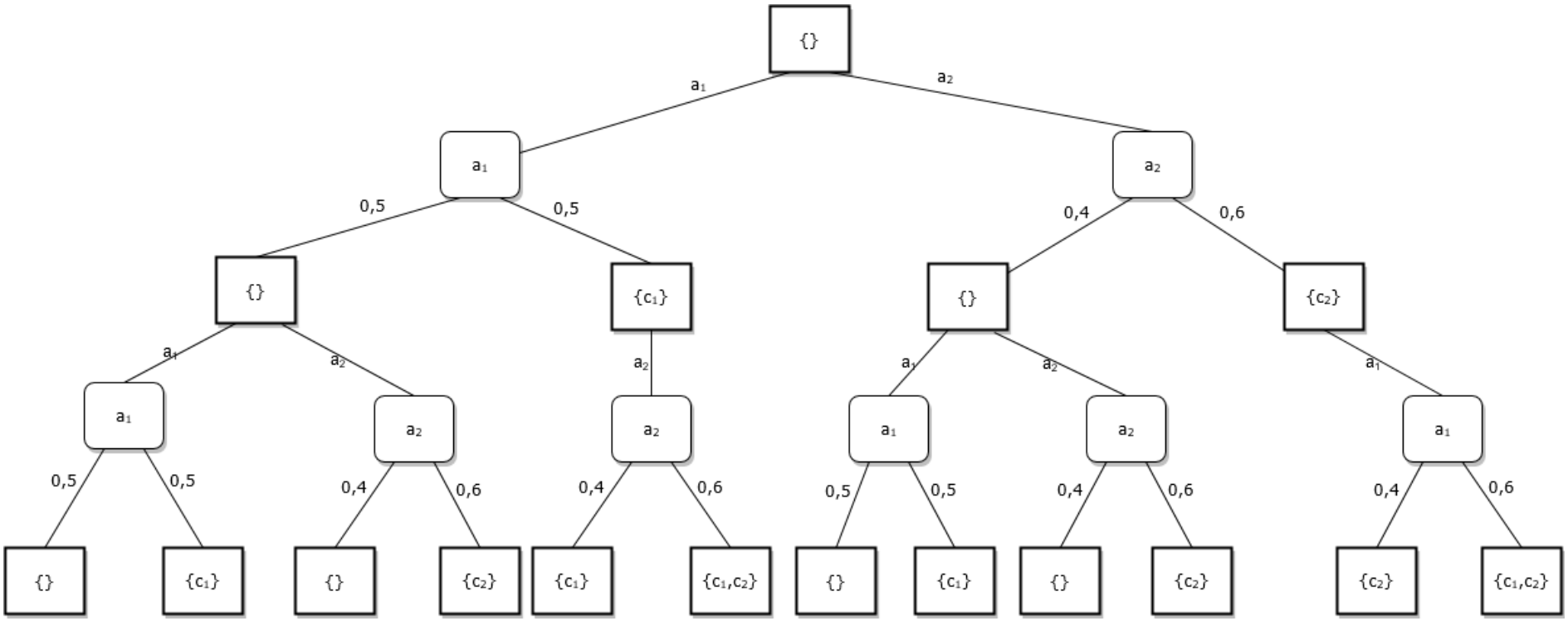


Figure: The corresponding tree for the academic advising domain

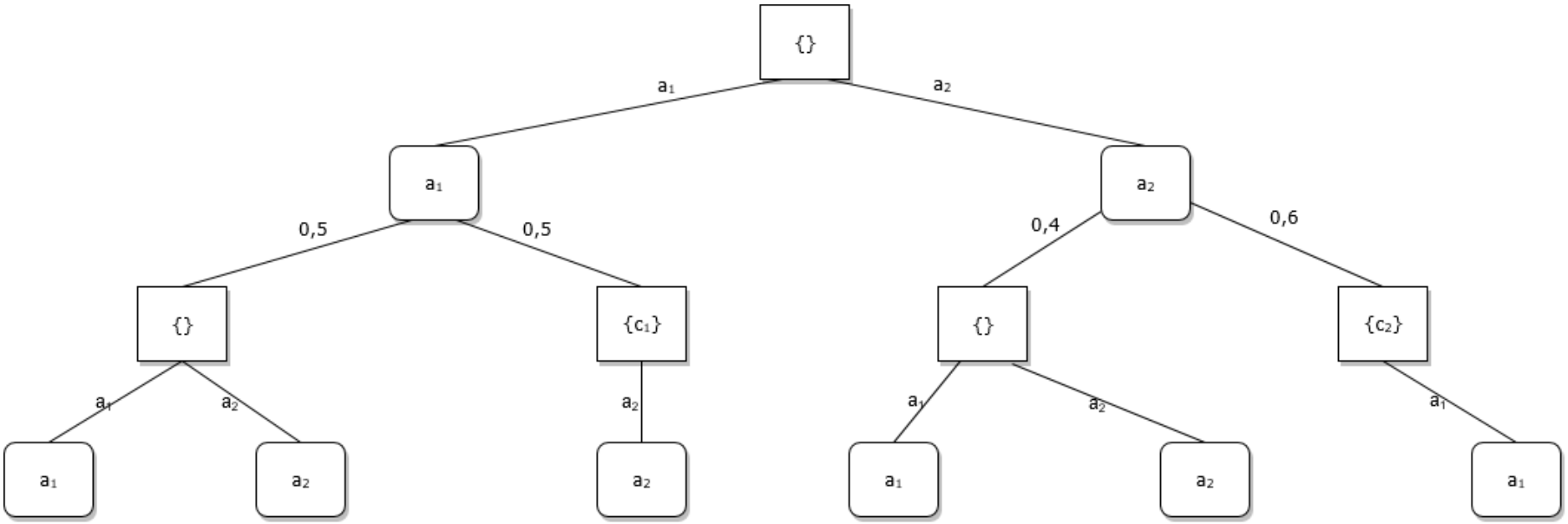
Upper confidence bounds applied to trees

build the search tree iteratively

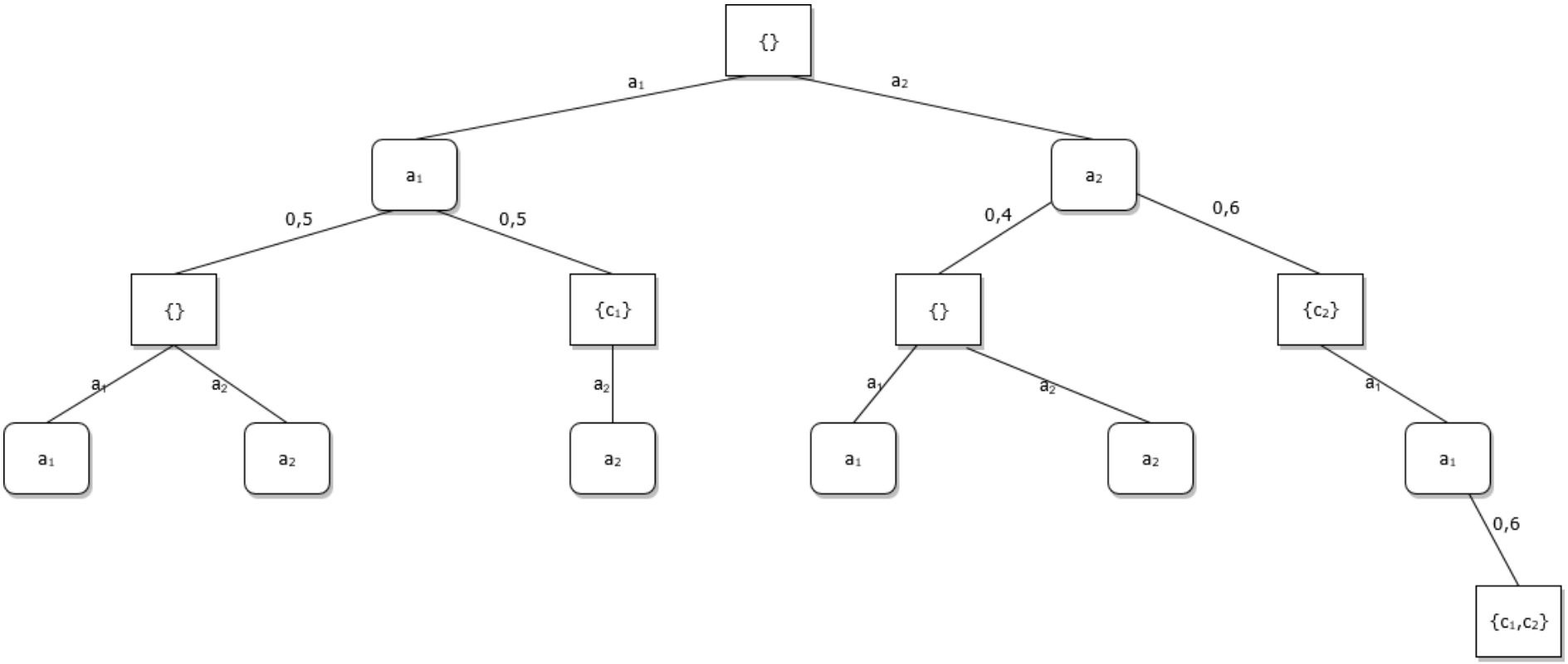
UCT phases:

- selection
- expansion
- backup

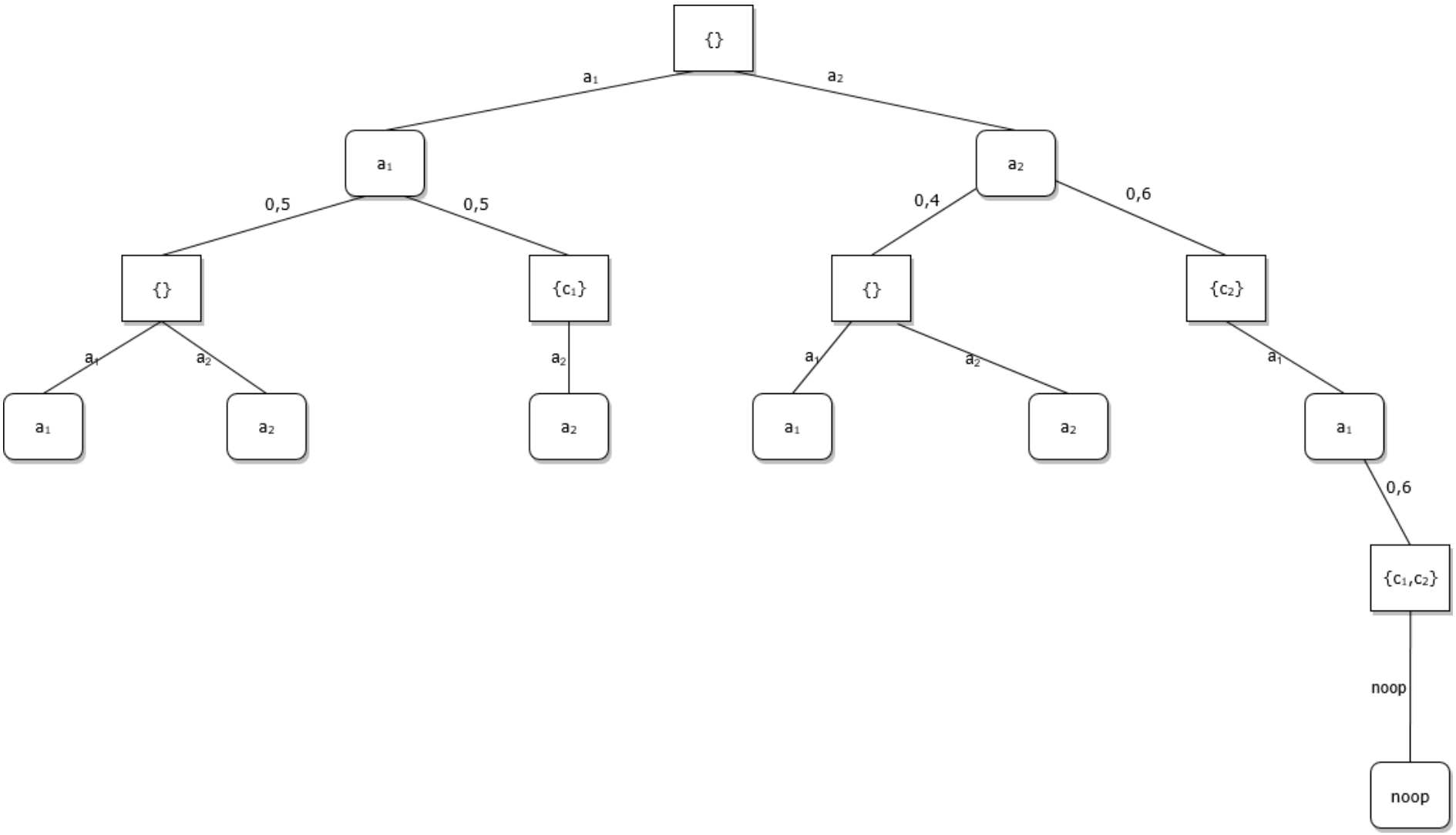
Selection phase



Expansion phase and Backup phase



Expansion phase and Backup phase



Equivalence relation

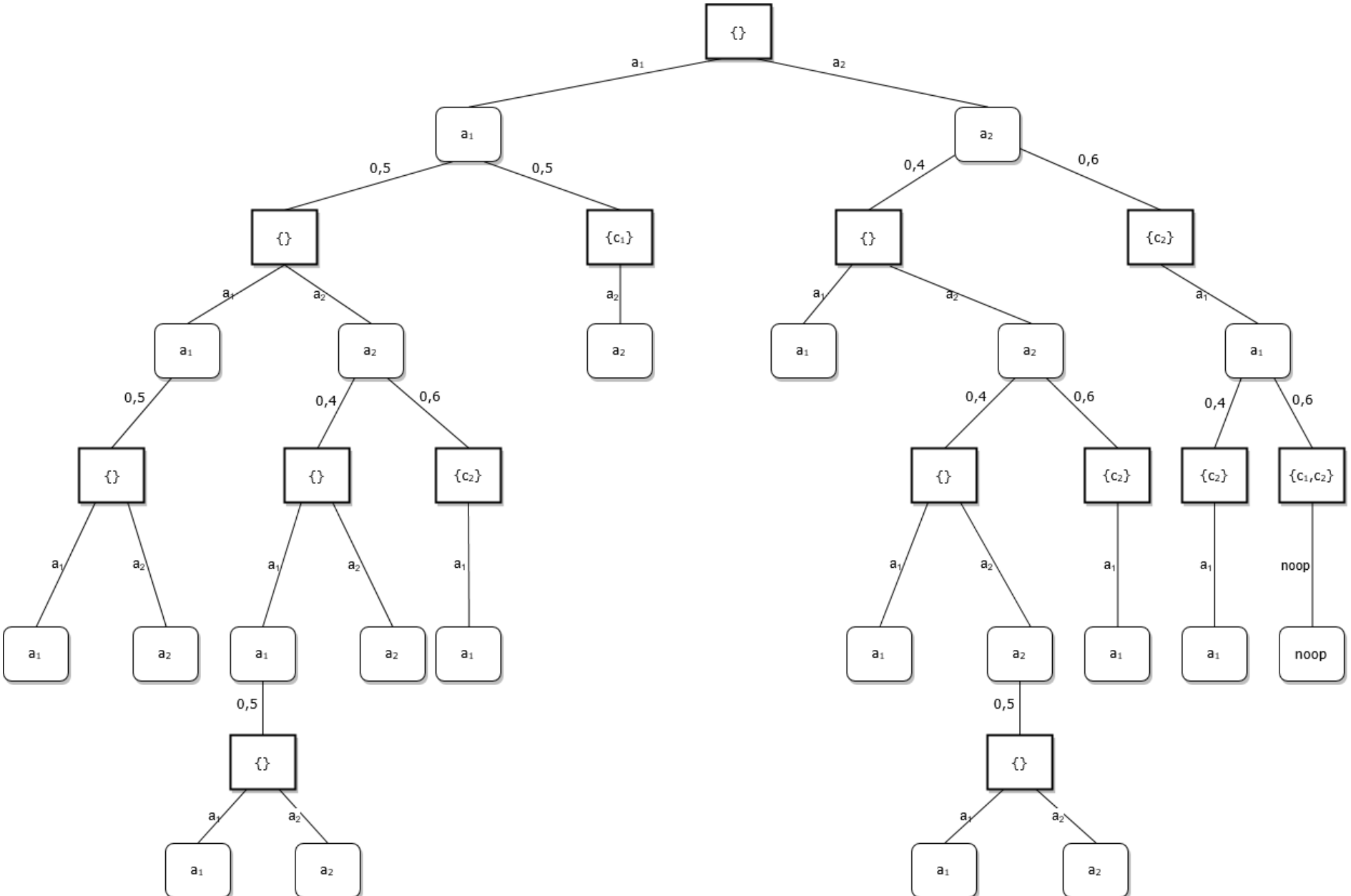
Definition

For two decision nodes $d_1 = \langle s_1, \hat{V}_1^t, N_1^t, h_1 \rangle$ and $d_2 = \langle s_2, \hat{V}_2^t, N_2^t, h_2 \rangle$, $d_1 \sim_d d_2$ if and only if $h_1 = h_2$, $\hat{V}_1^t = \hat{V}_2^t$ and for all $\langle d_1, c_1 \rangle \in E_c^t$ there is a $\langle d_2, c_2 \rangle \in E_c^t$ with $c_1 \sim_c c_2$ and vice versa.

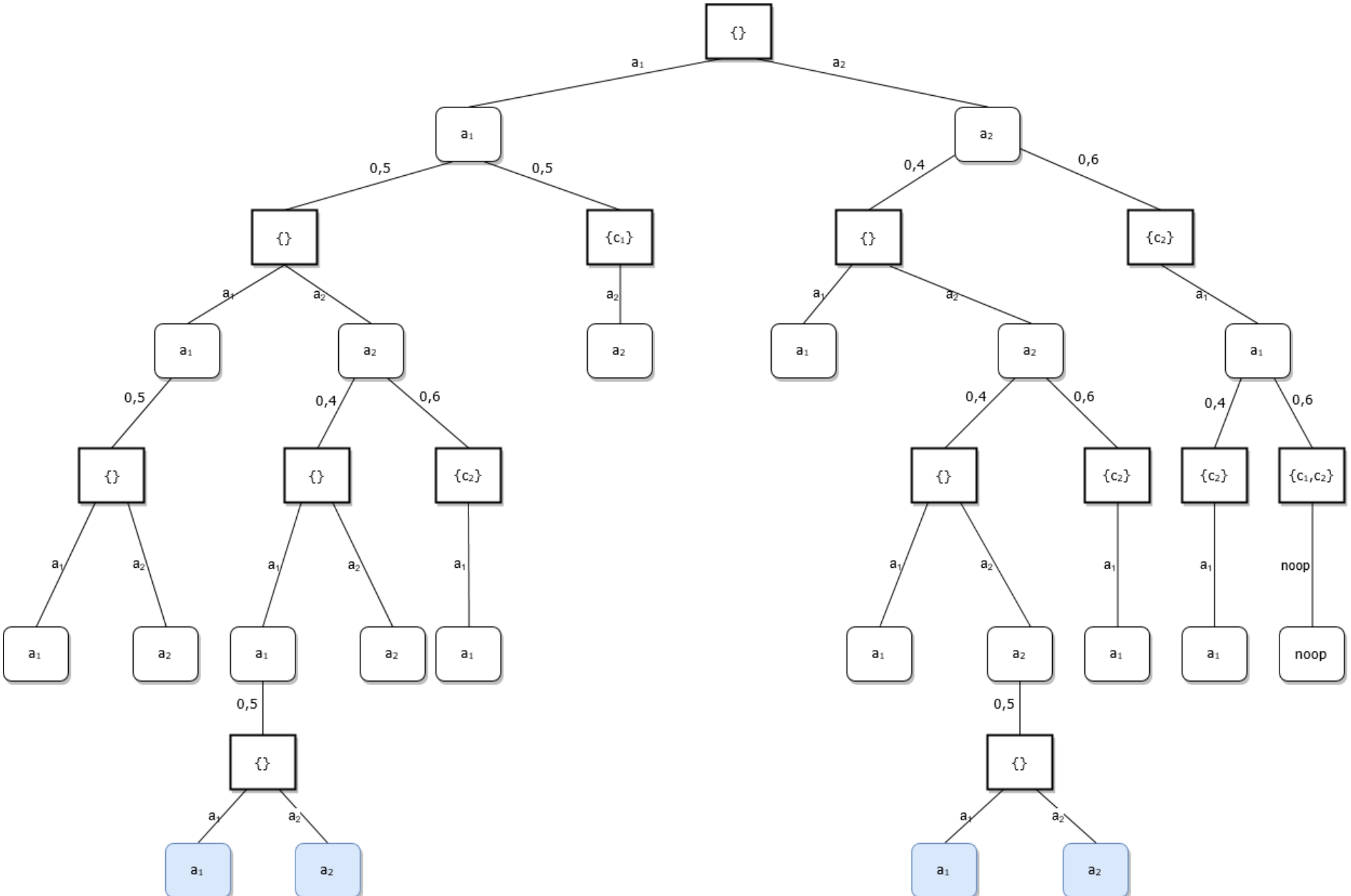
Definition

For two chance nodes $c_1 = \langle s_1, a_1, \hat{Q}_1^t, N_1^t, h_1 \rangle$ and $c_2 = \langle s_2, a_1, \hat{Q}_1^t, N_1^t, h_2 \rangle$ $c_1 \sim_c c_2$ if and only if $h_1 = h_2$, $\hat{Q}_1^t = \hat{Q}_2^t$ and for all $\langle c_1, d_1 \rangle \in E_c^t$ there is a $\langle c_2, d_2 \rangle \in E_d^t$ with $d_1 \sim_d d_2$ where $(L_{d_1} = L_{d_2})$ and vice versa.

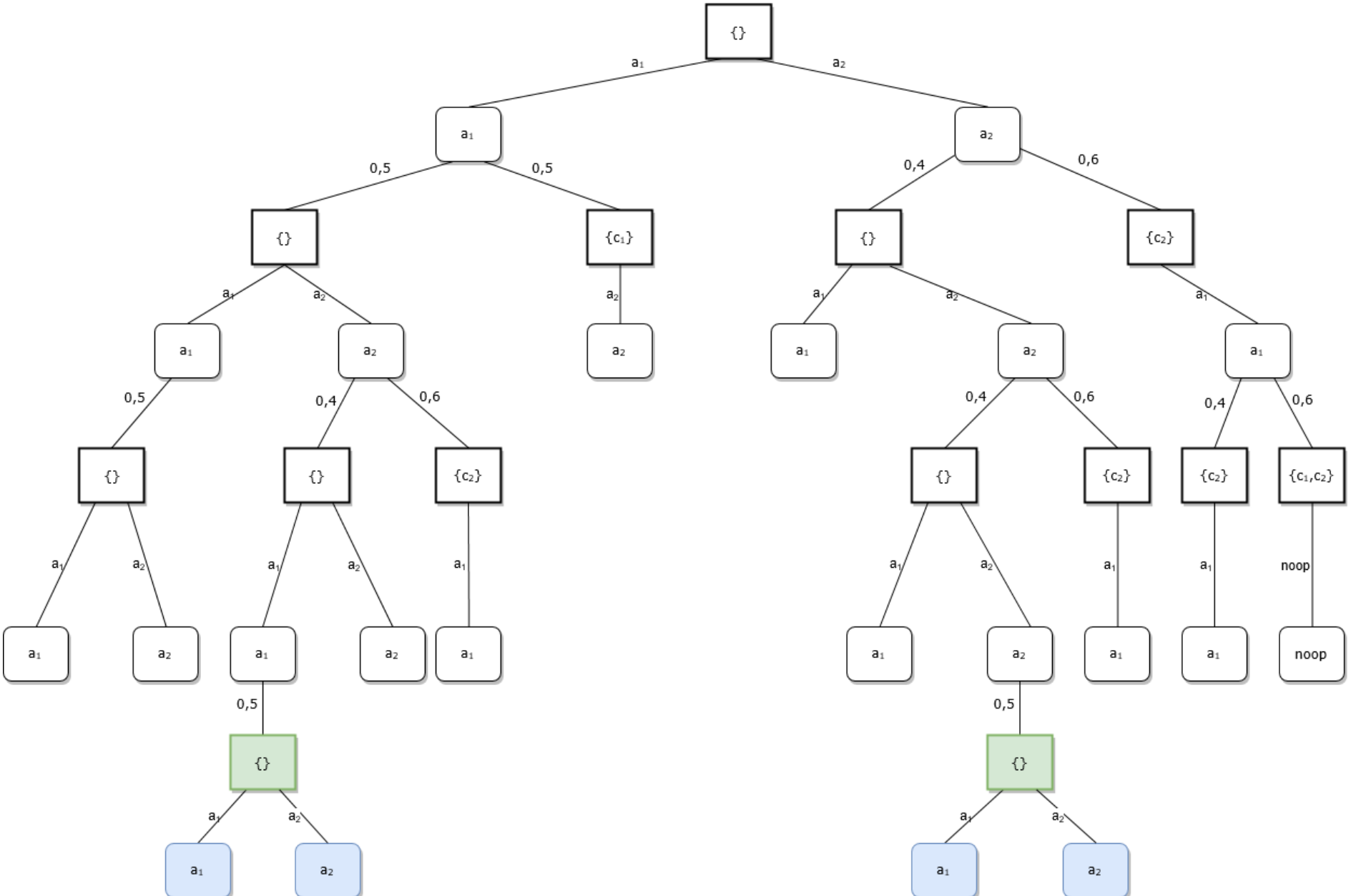
Example



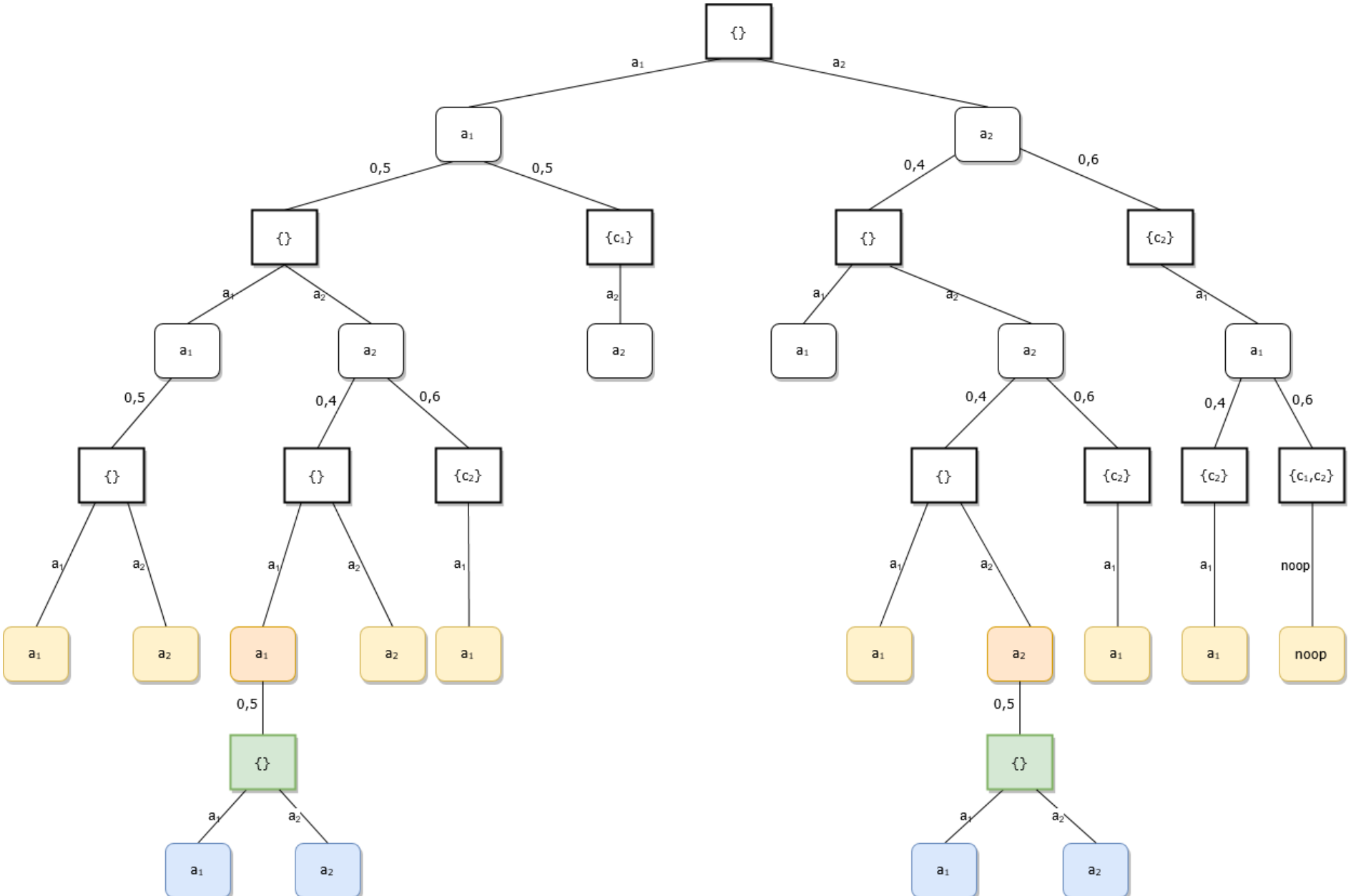
Example



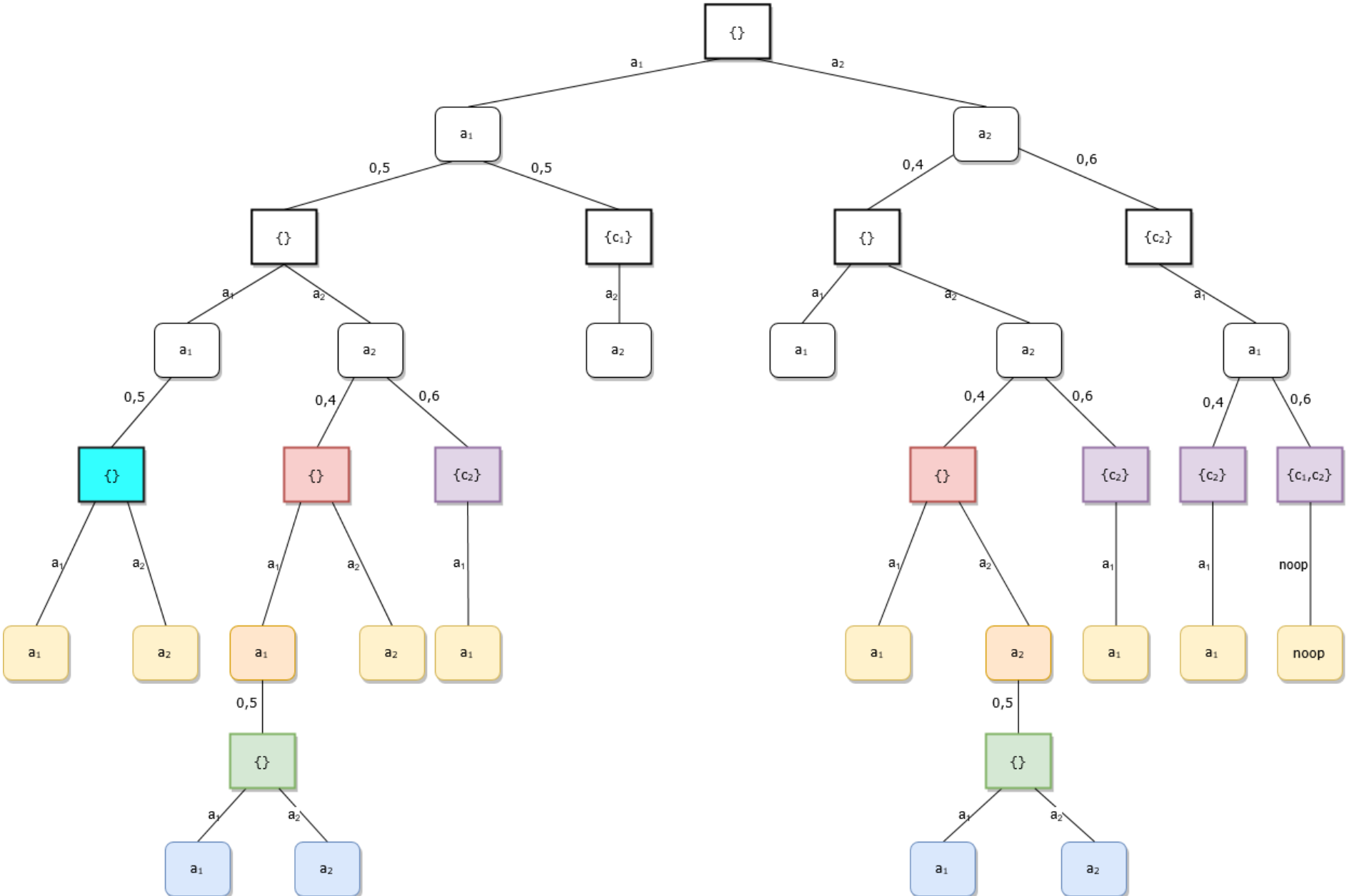
Example



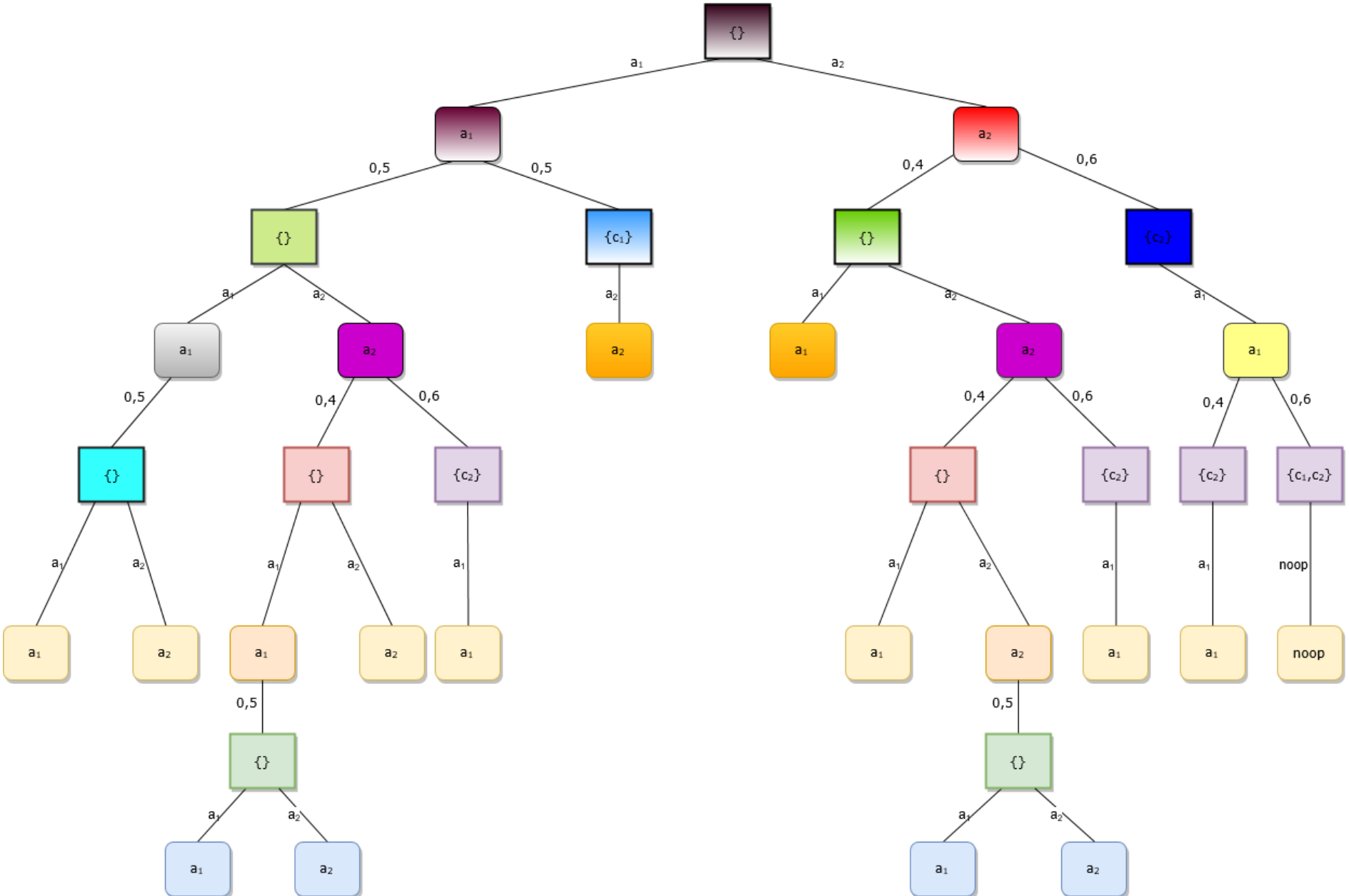
Example



Example



Example



Equivalence classes

Each equivalence class has a Q-value-mean M for $\bar{d}$ the mean of all the state-value-estimates

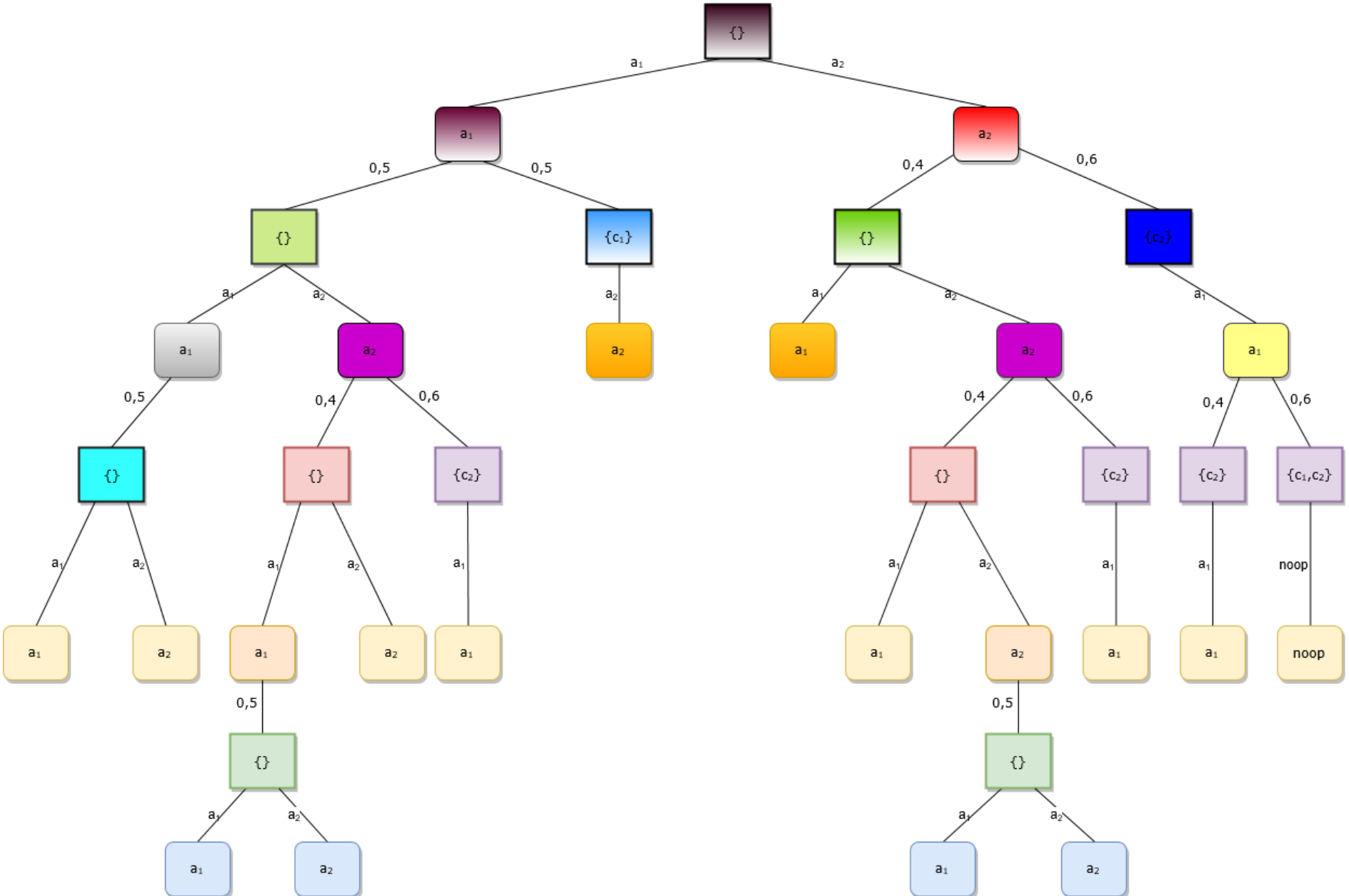
$$M(\bar{d}) = \frac{1}{|\bar{d}|} \sum_{d \in \bar{d}} \hat{Q}(d) \tag{1}$$

and for $\bar{c}$ the mean of all the action-value-estimates:

$$M(\bar{c}) = \frac{1}{|\bar{c}|} \sum_{c \in \bar{c}} \hat{V}(c) \tag{2}$$

- use the Q-value-mean in the selection phase

Example



Experiment

Framework

- C++
- UCT in the framework of Prost
- IPC benchmarks of 2011 and 2014

Parameter

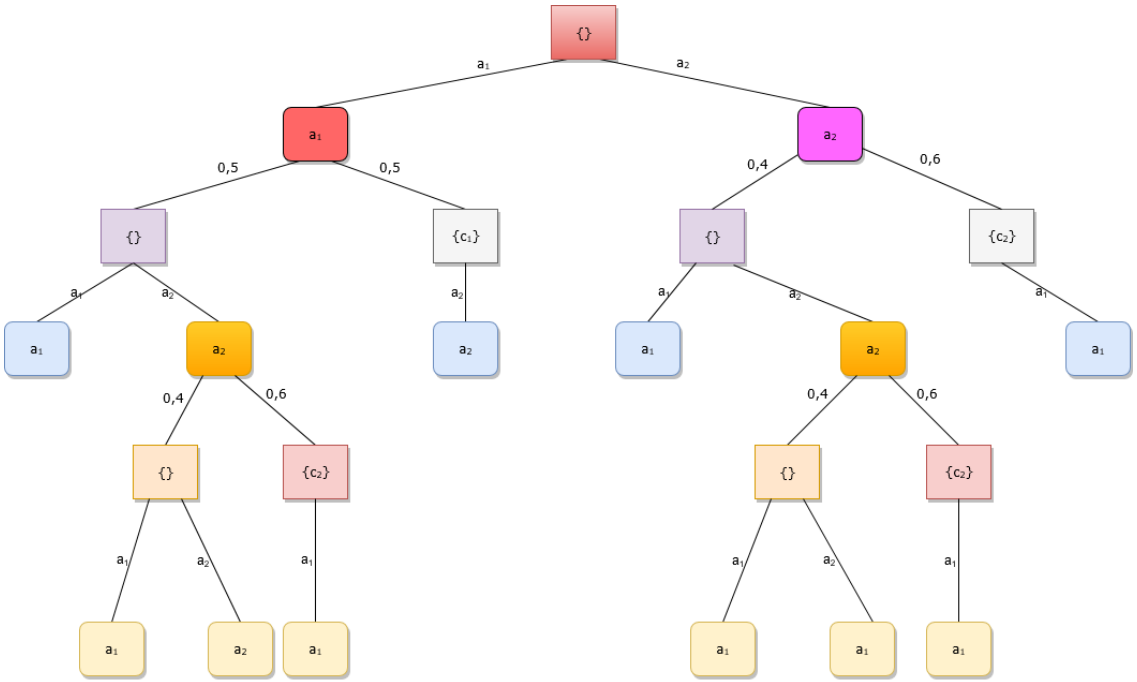
- update frequency τ
- trial length
- reasonable action
- heuristic: IDS or random walk
- expansion size

Results

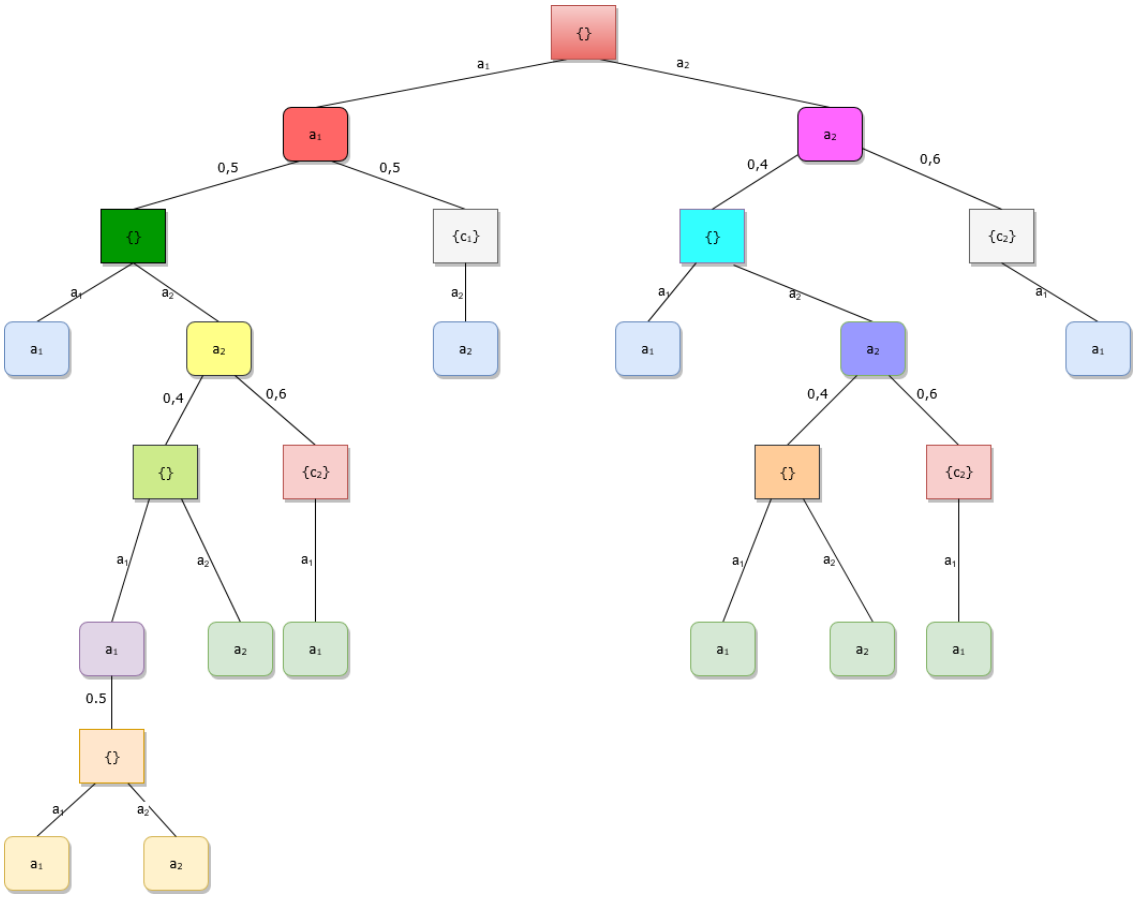
Results with activated reasonable action, IDS and the trial length of one

	wildfire	triangle	recon	elevators	tamarisk	sysadmin	academic	game	traffic	crossing	skill	navigation	Total
IDS-UCT^3	0.85	0.84	0.96	0.65	0.88	0.94	0.39	0.92	0.97	0.74	0.92	0.3	0.78
IDS-UCT^2	0.81	0.83	0.98	0.58	0.9	0.87	0.39	0.95	0.98	0.81	0.91	0.34	0.78
IDS-UCT^1	0.85	0.85	0.98	0.76	0.89	0.91	0.39	0.96	0.97	0.79	0.93	0.32	0.8
IDS-UCT^0	0.9	0.94	0.98	0.89	0.91	0.9	0.5	0.95	0.98	0.81	0.91	0.31	0.83

Problem of the abstraction



Problem of the abstraction



Summary

ASAP-UCT

- combines abstraction of states and of state-action pairs with UCT
- generates equivalence classes based on the equivalence classes of the children
- flaw in the abstraction in incomplete trees

References

Ankit Anand et al. (2015)

ASAP-UCT: Abstraction of State-Action Pairs in UCT

Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI,1509–1515

.

Levente Kocsis and Csaba Szepesvári, (2006)

Bandit Based Monte-Carlo Planning

Proceedings of the 17th European Conference on Machine Learning, ECML,282–293.

Martin Puterman (1994)

Markov Decision Processes: Discrete Stochastic Programming

Wiley

UCB1

Exploration and Exploitation

$$c = \underset{\langle d,c\rangle \in E_c}{argmax}(Q^t(c) + V^t(d)\sqrt{\frac{N^t(d)}{N^t(c)}})$$