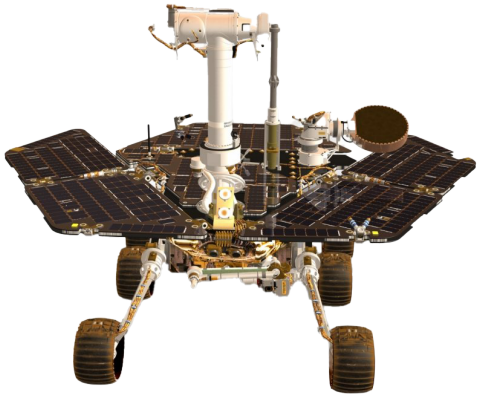
SOGBOFA as heuristic guidance for THTS

Ferdinand Badenberg

Universität Basel

20.5.2020

Problem Setting



Problems based on real life problems, such as:

- Academic Advising
  - Students take courses to graduate
  - Probability to pass a course higher if prerequisite courses were passed
- Cooperative Recon
  - Mars rovers looking for life
  - Working together leads to a higher probability of success.

Markov Decision Process

The probabilistic planning problem is given as a Markov Decision Process with:

- A finite set of state variables inducing the states
- An initial state
- A finite set of action variables inducing the actions
- A transition function (over the state and action variables) for each state variable, modelling the probability of that variable being true in the next state, e.g. $s_0' = s_2 \wedge a_2$.
- A reward function over the state and action variables
- A finite horizon

Encoded as a RDDL task.

Monte-Carlo Tree Search

Build a search tree over trials:

1. Selection: Sample trajectories of actions following a tree policy
2. Expansion: Add new node(s), alternating between decision nodes ($\approx$ states) and chance nodes ($\approx$ actions)
3. Simulation: Initialize new node with a heuristic value
4. Backpropagation: Update the tree with the new information

Tree with branches for each action choice and each action outcome.

Other ways to provide a good estimate with very few samples?

SOGBOFA

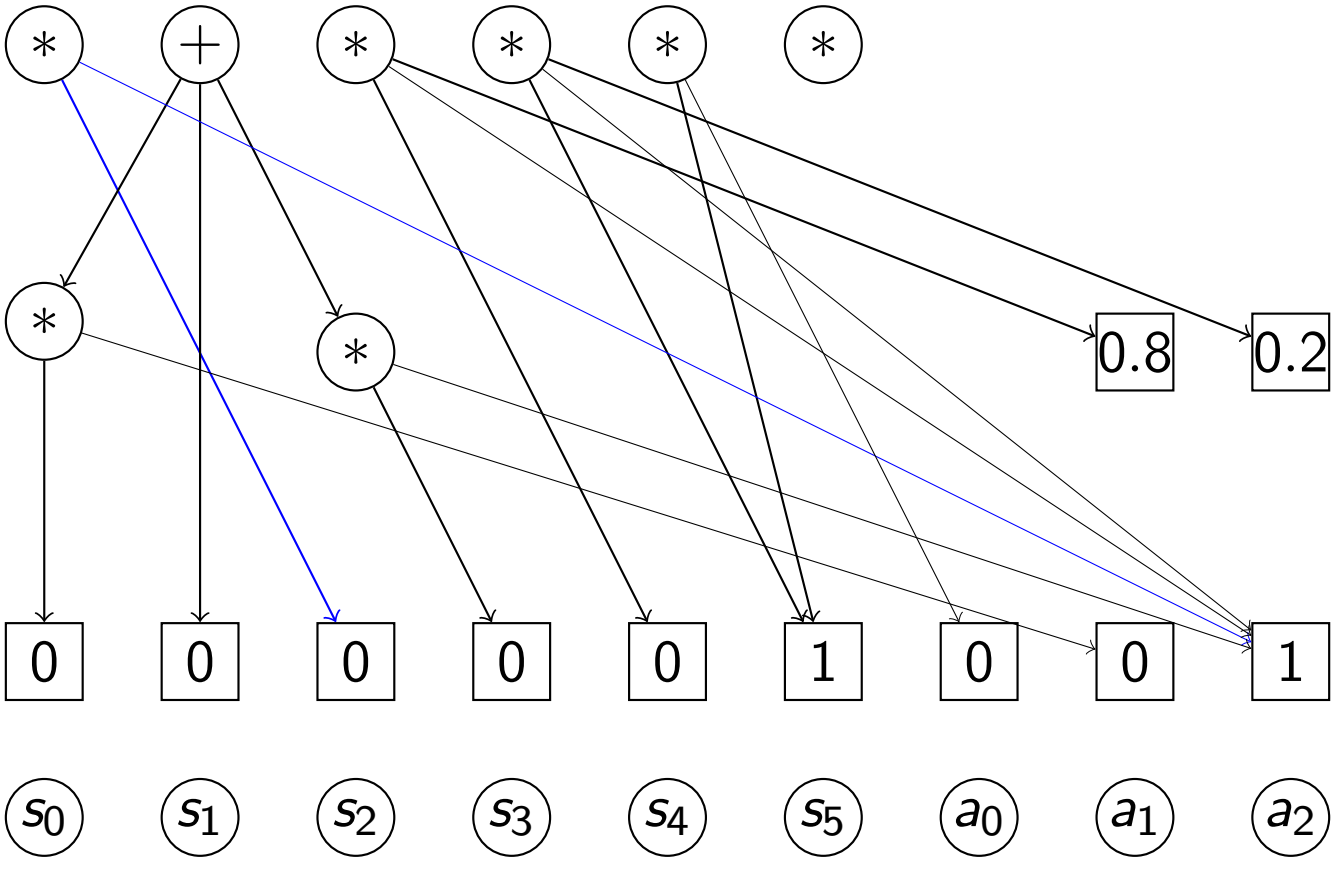
- Aggregating states
- Simplification: independence assumption of actions and states
- Eliminate branching for actions and outcomes!
- Loose asymptotic optimality
- Estimate long term reward as an algebraic function with actions as input

SOGBOFA Graph

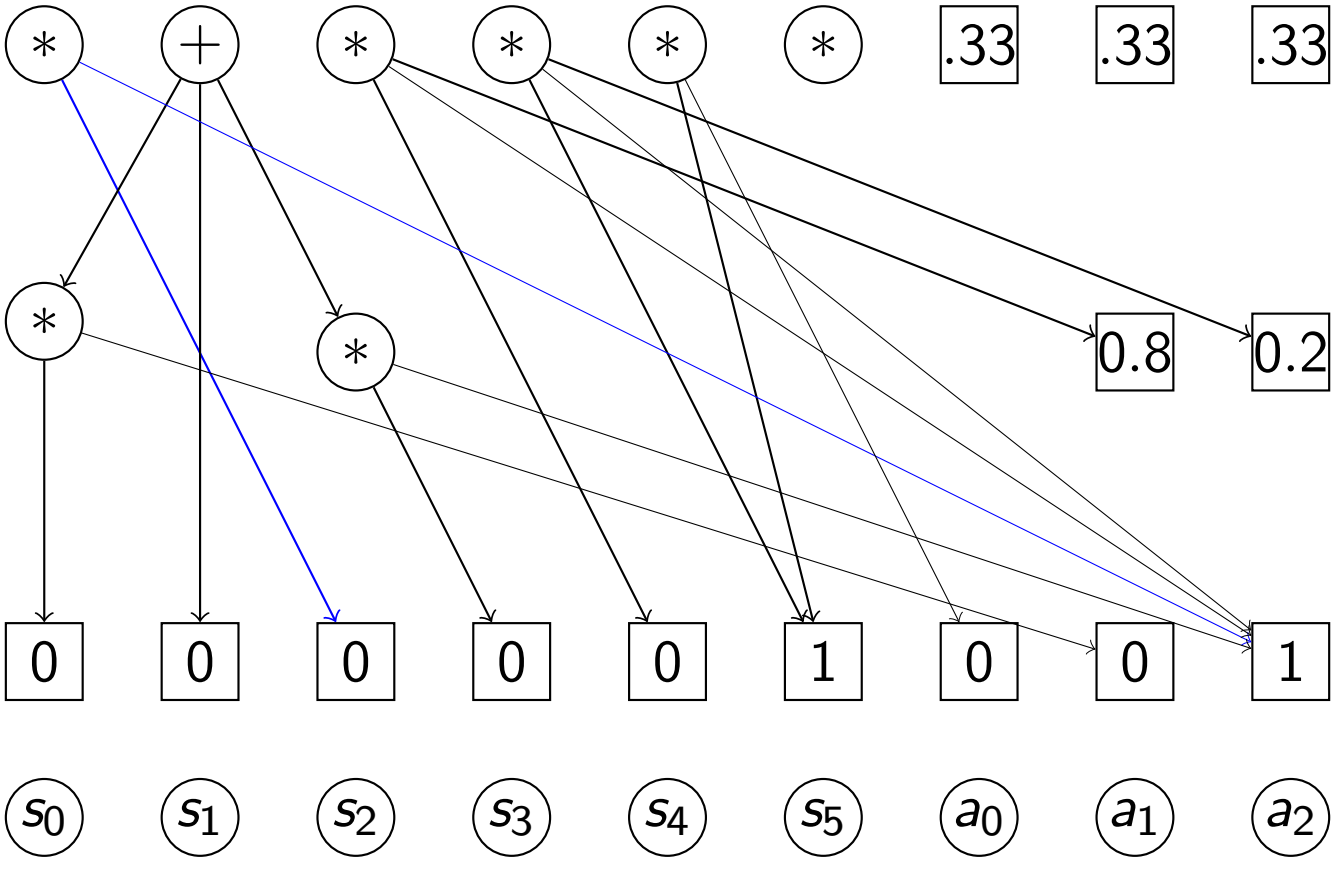
How can we represent the Q value as a function based on the action inputs?

1. RDDL description of the MDP describing the planning task
2. Convert RDDL expressions to arithmetic expressions (e.g. $s_0' = s_2 \wedge a_2$ becomes $s_0' = s_2 \cdot a_2$)
3. Build a graph over multiple steps using arithmetic expressions

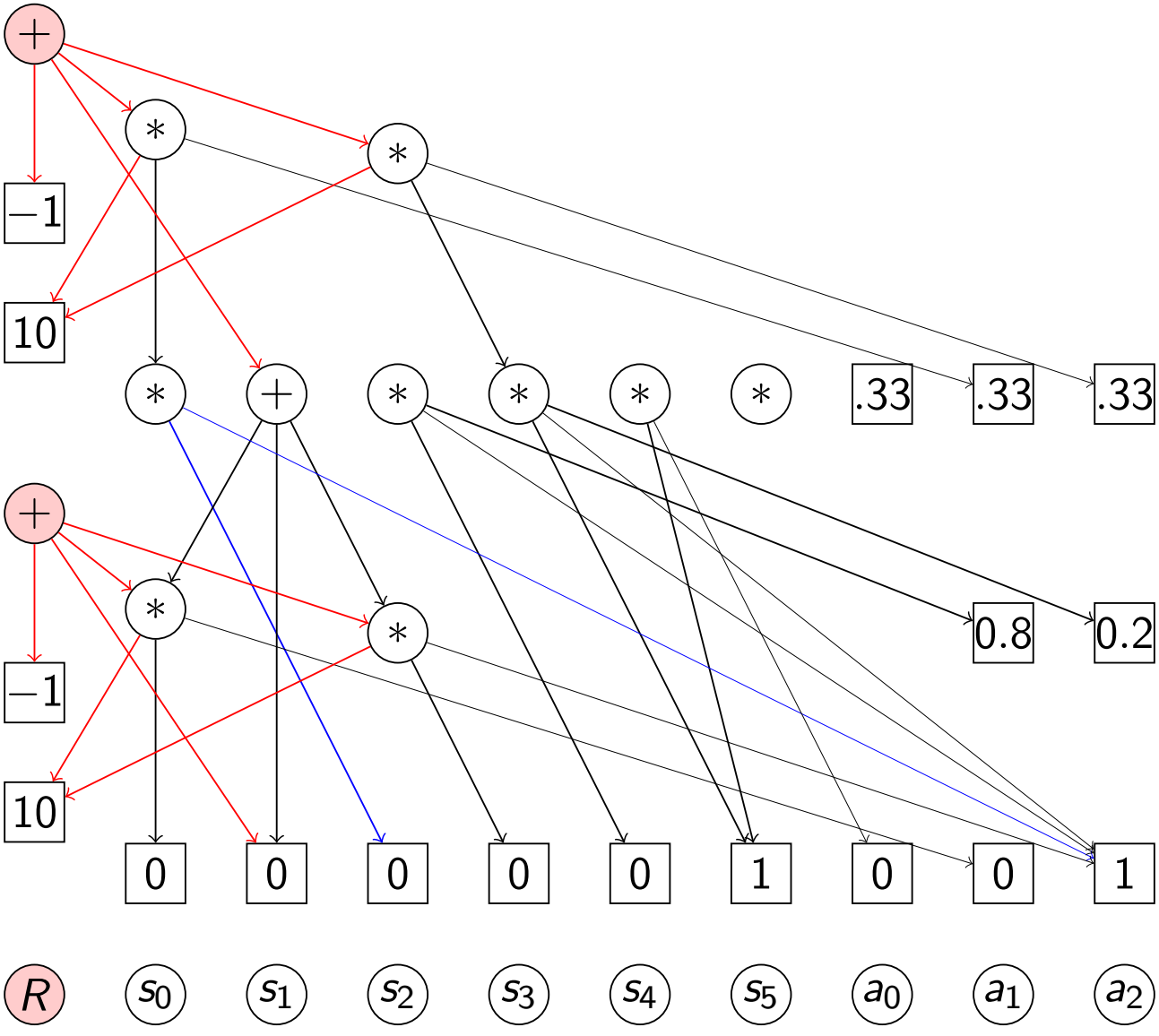
SOGBOFA Graph



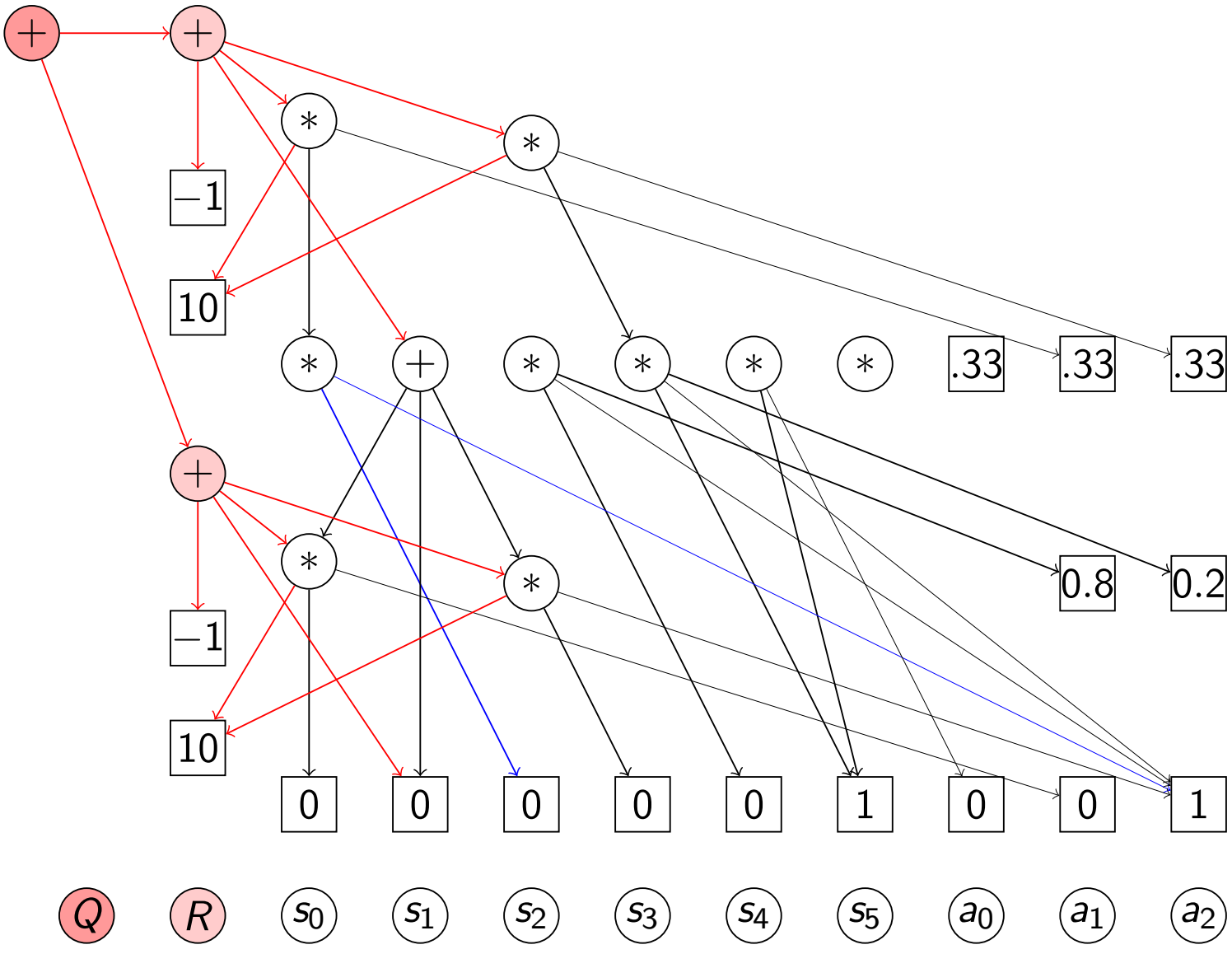
SOGBOFA Graph



SOGBOFA Graph



SOGBOFA Graph



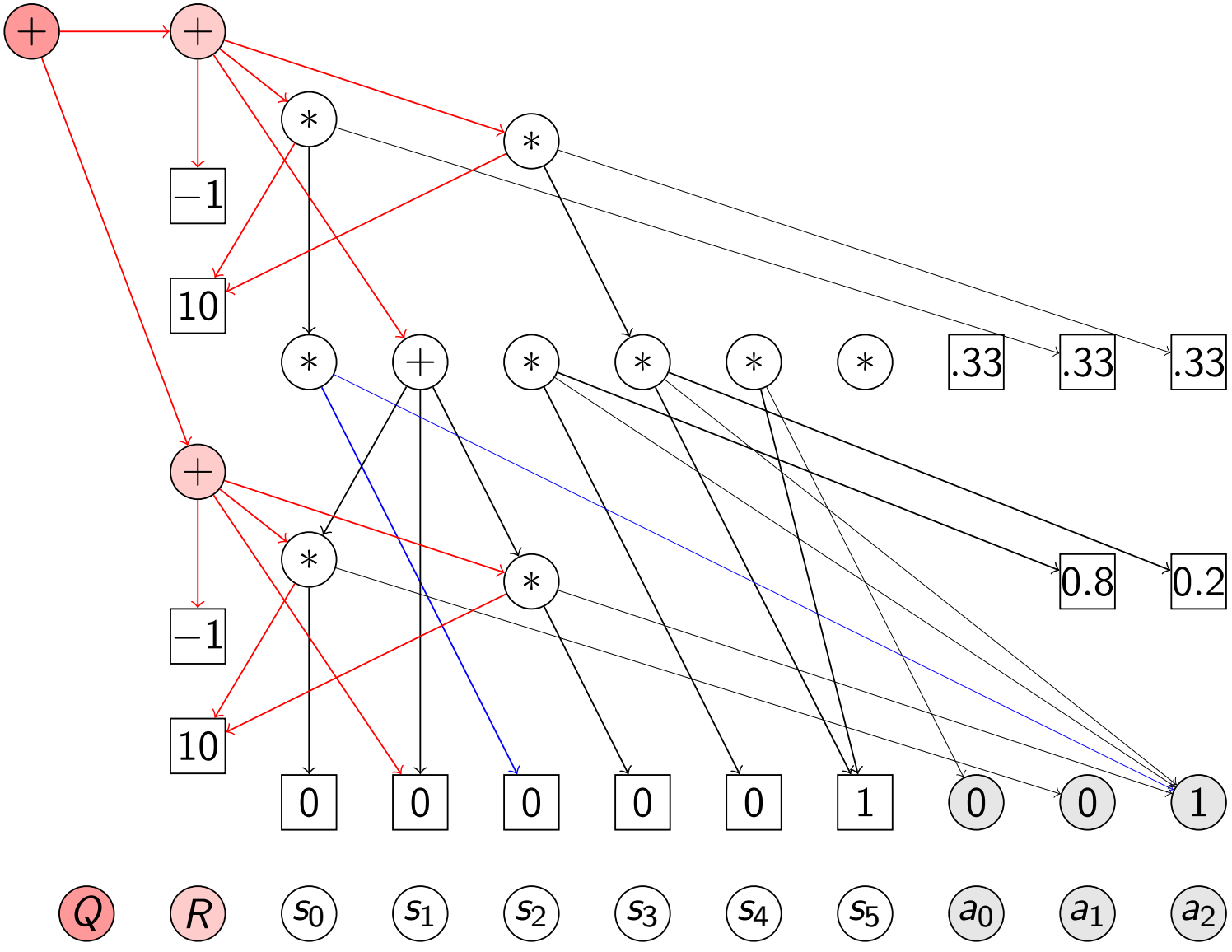
SOGBOFA: Notes

- The graph scales linearly with the simulated planning steps
- All information on dependence between the different actions and states is disregarded
- Marginal probabilities are still accurate

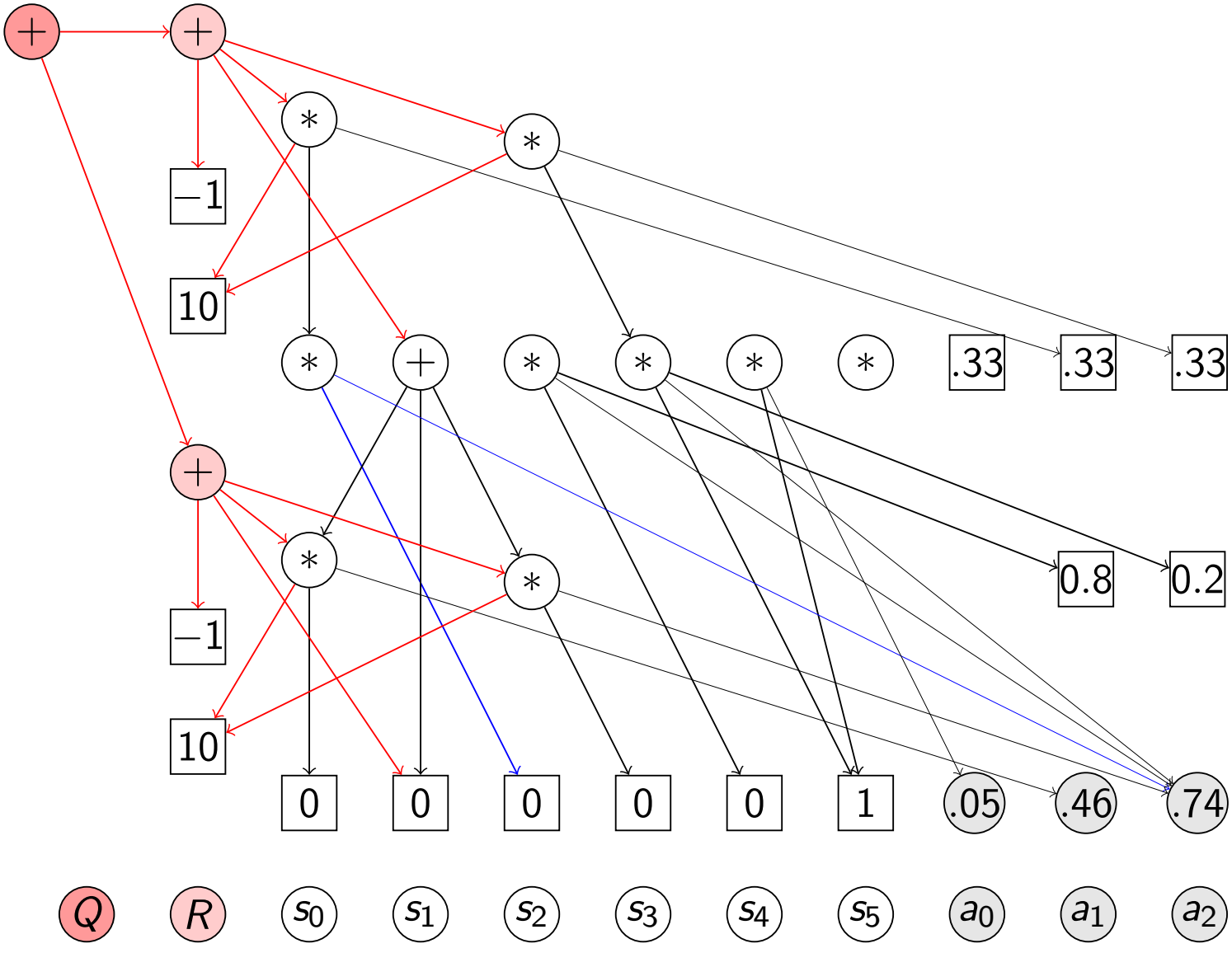
Optimizing Initial Actions

- Given: Differentiable Q value functions with our current actions as input
- Actions can be optimized with gradient ascent!
- Pick a random starting action state. Optimize it by repeating gradient ascent steps.

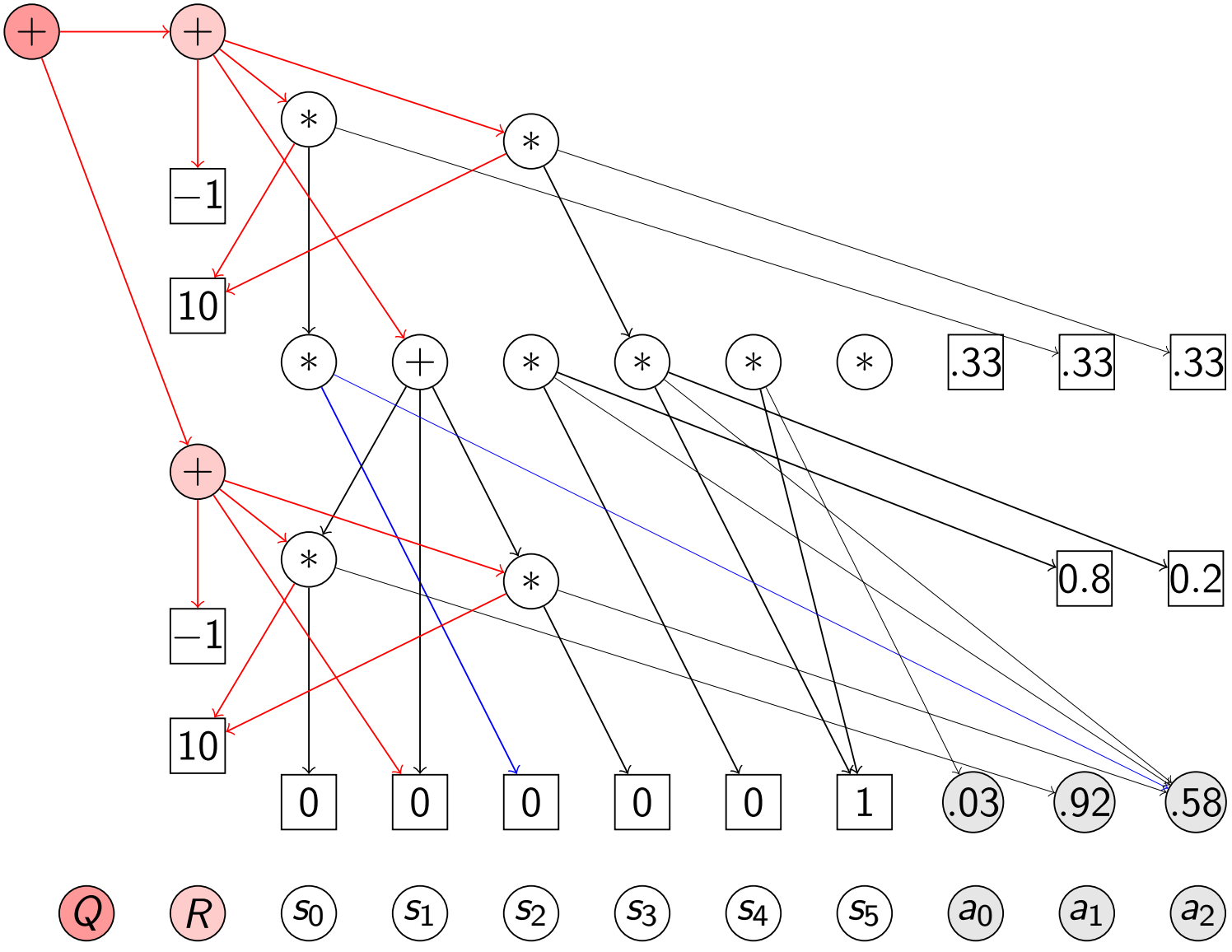
SOGBOFA Graph: Optimizing Initial Actions



SOGBOFA Graph: Optimizing Initial Actions



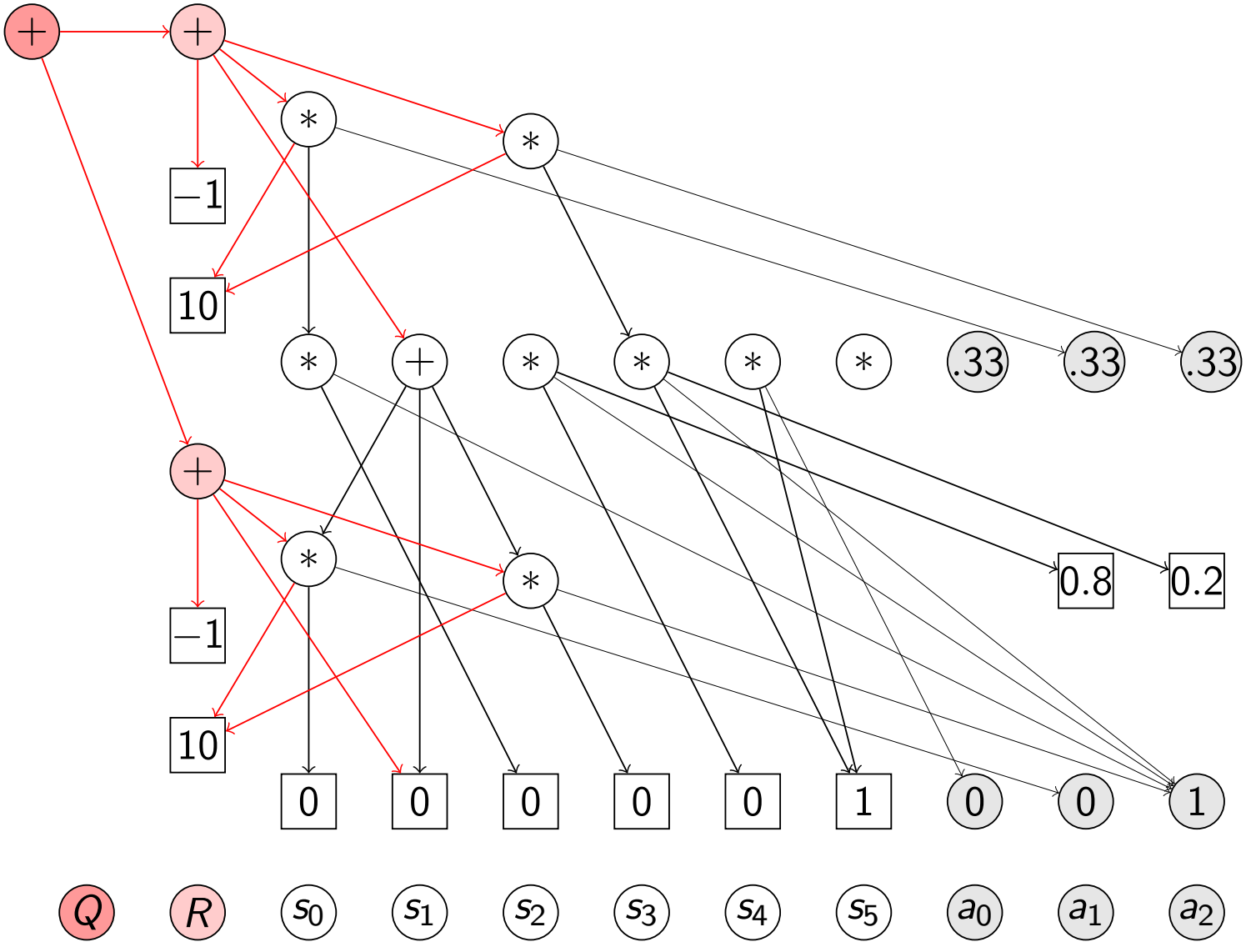
SOGBOFA Graph: Optimizing Initial Actions



Optimizing Future Actions

- Future actions are very uninformative ($\approx$ random policy)
- Conformant SOGBOFA algorithm also optimizes future actions
- With reverse mode automatic differentiation, the full gradient can be calculated in a single traversal of the graph

SOGBOFA Graph: Optimizing Future Actions



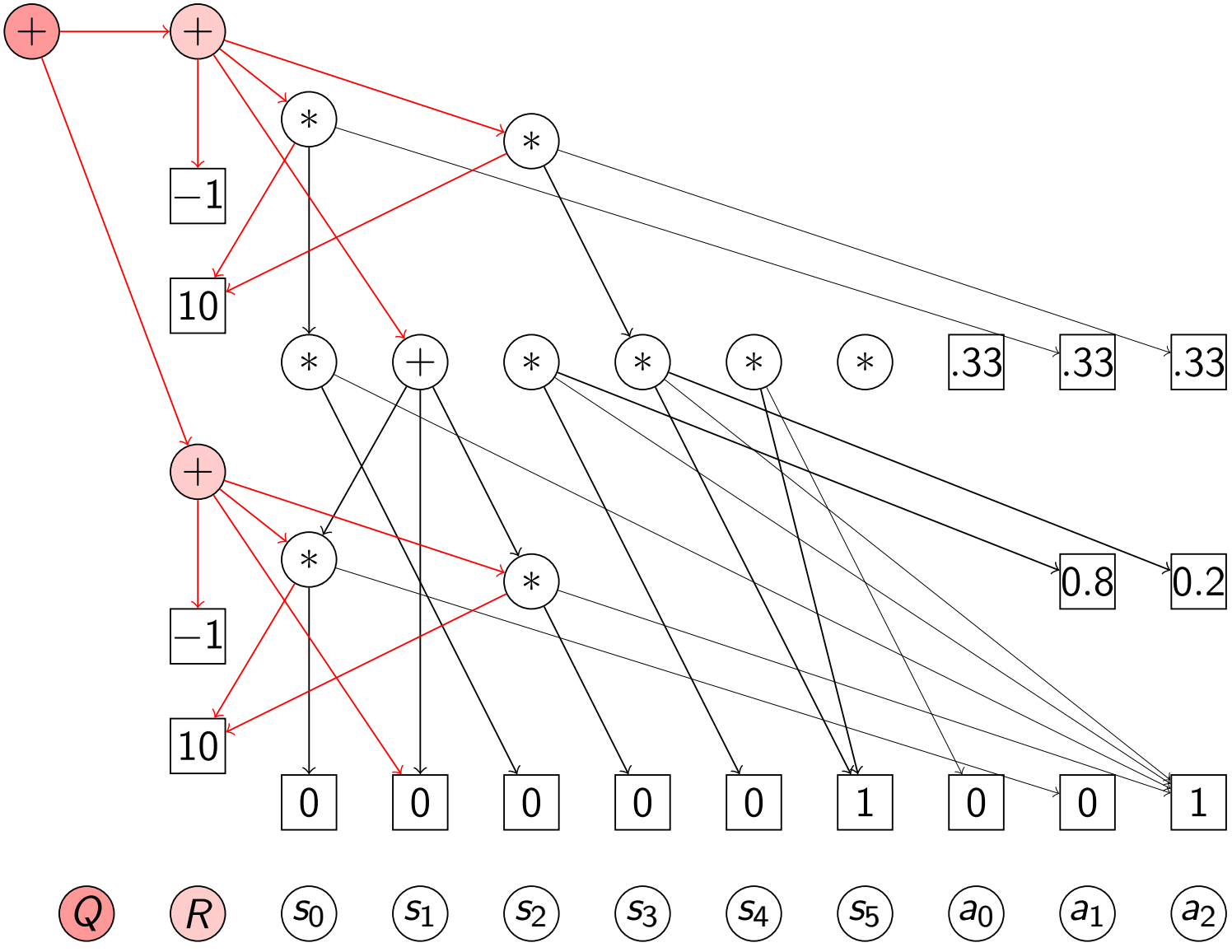
Heuristics from SOGBOFA

- Before: Optimize the actions to find the best actions in the current state
- Now: Evaluate the quality of given actions in the current state
- Actions at the input level are now fixed

Propagation Heuristic

- Estimate the Q values in a single forward propagation of the action values through the SOGBOFA graph.
- Uses uniform values for future actions
- No gradient steps or optimization of actions

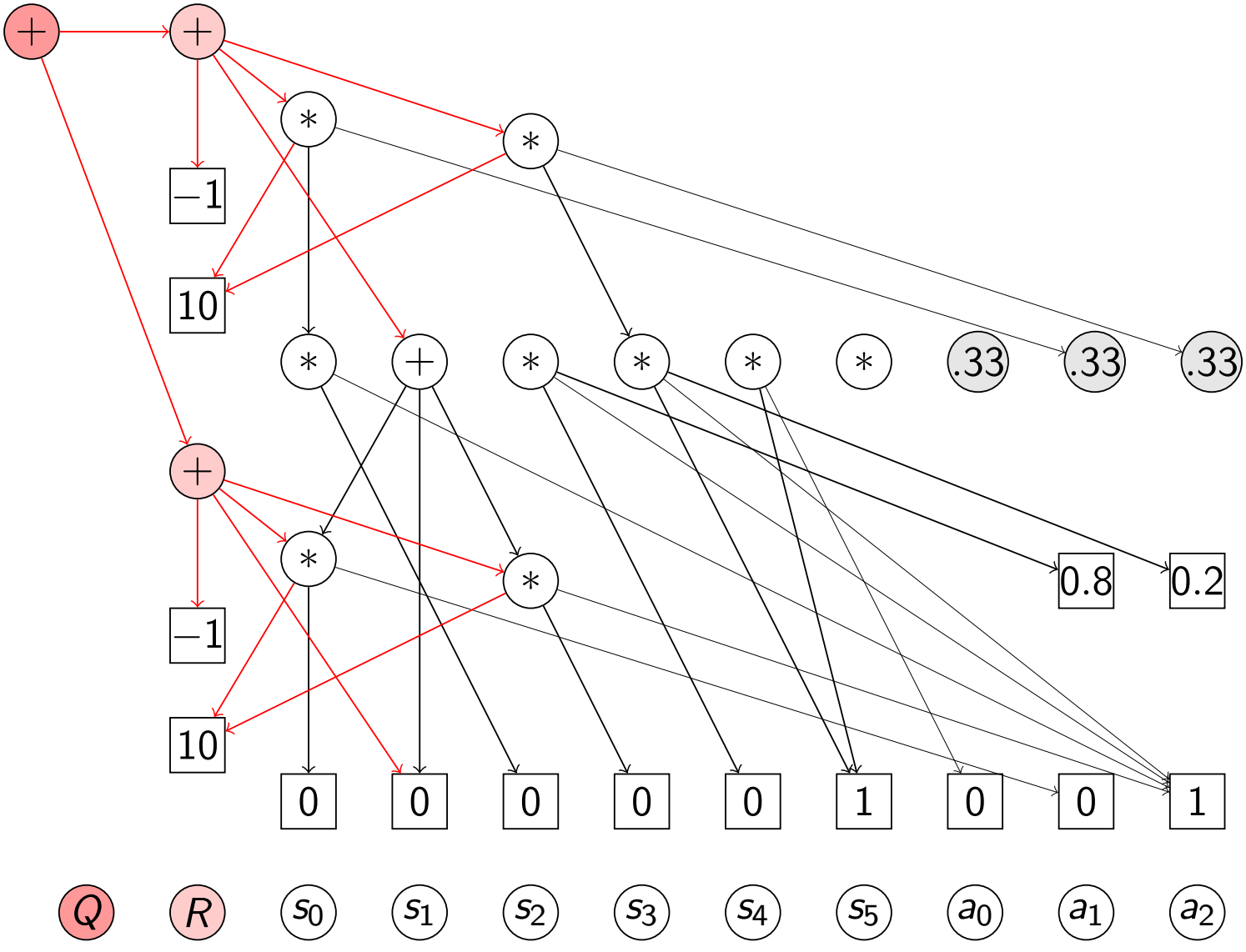
Propagation Heuristic SOGBOFA Graph



Conformant Heuristic

- Motivation: Include gradient-based optimization
- Optimize the future actions over few gradient steps
- Estimate the Q values as the evaluation of the SOGBOFA graph with the optimized actions
- Better guidance through optimized future actions, but slower

Conformant Heuristic SOGBOFA Graph



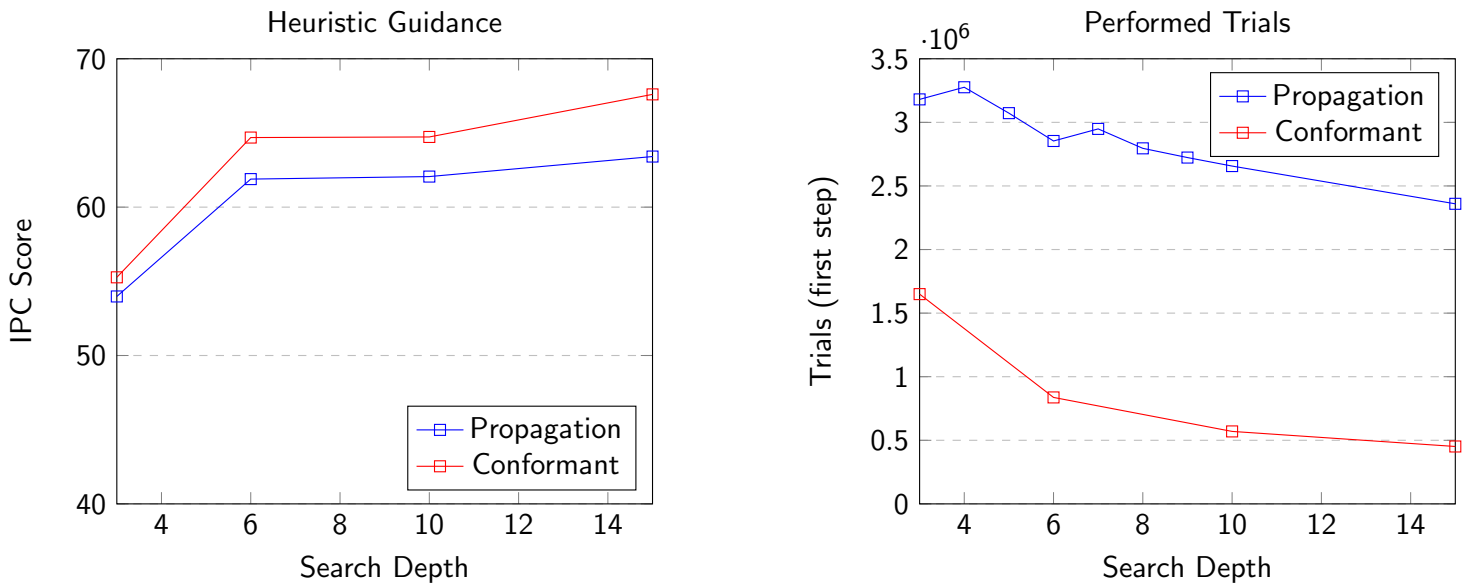
Evaluation

- Online planning setting: alternate planning and action execution
- Comparison to PROST IPC2014 with the IDS heuristic.

Parameter: Search Depth

How many future steps should we consider?

Figure: Search Depth affecting Heuristic Guidance and Calculation Time



Why is the conformant heuristic so much slower?

Performance: Overview

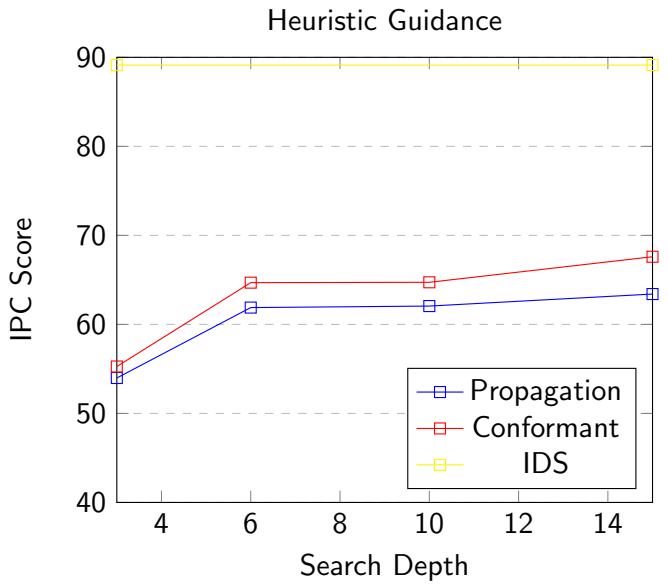
Table: IPC Scores for both Heuristic (respective best Configurations)

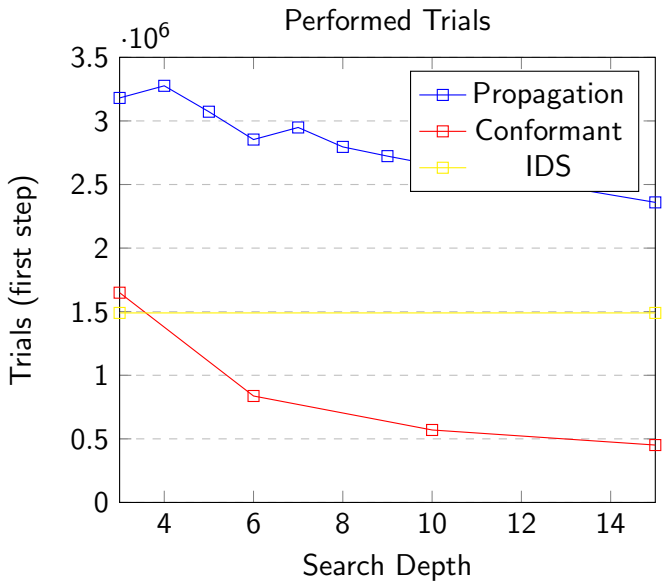
Domain	Propagation Heuristic	Conformant Heuristic
crossing-traffic-2011	**9.72**	8.07
elevators-2011	9.28	**9.55**
game-of-life-2011	**9.02**	8.57
navigation-2011	**9.31**	9.28
recon-2011	9.57	**9.61**
skill-teaching-2011	9.09	**9.30**
sysadmin-2011	**7.45**	5.76
academic-advising-2014	**3.61**	3.06
tamarisk-2014	**9.65**	7.52
triangle-tireworld-2014	**6.37**	4.92
wildfire-2014	**8.99**	8.59
academic-advising-2018	**4.72**	3.62
cooperative-recon-2018	**10.23**	3.96
Sum	**107.00**	91.81

Evaluation: Comparison to IDS

How does this compare to IDS from PROST IPC2014?

Figure: Heuristic Guidance and Calculation Time Compared to IDS





Performance: Comparison to IDS

Table: IPC Scores for both Heuristic (respective best Configurations) against IPC2014

Domain	Prost IPC2014	Propagation Heuristic	Conformant Heuristic
crossing-traffic-2011	8.66	**9.72**	8.07
elevators-2011	9.38	9.28	**9.55**
game-of-life-2011	**9.60**	9.02	8.57
navigation-2011	8.88	**9.31**	9.28
recon-2011	9.52	9.57	**9.61**
skill-teaching-2011	9.07	9.09	**9.30**
sysadmin-2011	6.76	**7.45**	5.76
academic-advising-2014	2.99	**3.61**	3.06
tamarisk-2014	7.64	**9.65**	7.52
triangle-tireworld-2014	**7.61**	6.37	4.92
wildfire-2014	5.52	**8.99**	8.59
academic-advising-2018	3.23	**4.72**	3.62
cooperative-recon-2018	9.58	**10.23**	3.96
Sum	98.44	**107.00**	91.81

Conclusion

- The propagation heuristic is very fast to calculate, yet reasonably informative.
- The SOGBOFA graph can lead to strong results when used as heuristic guidance for THTS.
- The conformant heuristic is better informed, but suffers from limited trials.
- A custom implementation of gradient calculation would significantly improve the performance of the conformant heuristic.

Questions?

Thank You!

Action Constraints

- Important information through action constraints is lost
- Sum constraints on actions $\sum a_i \leq B$ are supported
- Added through projection of actions to satisfy constraints
- More general way to add any action constraint from action preconditions?
- Observation: All preconditions are algebraic formulas
- Idea: integrate them into graph by adding a penalty to the reward for violated action preconditions

Evaluation: Overview

Table: IPC Scores for both Versions of the Standalone Planner and Heuristic (respective best Configurations) against IPC2014

Domain	Prost	Planner	C. Planner	Propagation	Conformant
crossing-traffic-2011	8.66	4.19	4.19	**9.72**	8.07
elevators-2011	9.38	0.04	0.04	9.28	**9.55**
game-of-life-2011	9.60	4.86	4.79	9.02	8.57
navigation-2011	8.88	0.24	0.24	9.31	9.28
recon-2011	9.52	0.00	0.00	9.57	**9.61**
skill-teaching-2011	9.07	8.39	8.02	9.09	**9.30**
sysadmin-2011	6.76	**9.70**	**9.75**	7.45	5.76
academic-advising-2014	2.99	1.18	0.00	3.61	3.06
tamarisk-2014	7.64	6.37	6.08	**9.65**	7.52
triangle-tireworld-2014	7.61	1.08	1.09	6.37	4.92
wildfire-2014	5.52	**9.68**	**9.70**	8.99	8.59
academic-advising-2018	3.23	**6.68**	**4.76**	4.72	3.62
cooperative-recon-2018	9.58	1.79	0.94	**10.23**	3.96
Sum	98.44	54.17	49.58	**107.00**	91.81

Evaluation: Standalone

Table: Effect of Generalized Action Constraints on the IPC score

Domain	Generalized	Sum	Generalized Conformant	Sum Conformant
crossing-traffic-2011	9.83	9.79	9.81	9.59
elevators-2011	0.29	0.29	5.82	3.77
game-of-life-2011	6.86	8.52	7.59	8.07
navigation-2011	2.89	2.89	4.79	4.00
recon-2011	0.00	0.00	0.00	0.00
skill-teaching-2011	8.94	9.19	6.28	8.96
sysadmin-2011	8.39	9.75	8.45	8.82
academic-advising-2014	1.23	1.23	0.00	0.00
tamarisk-2014	9.19	9.27	5.39	8.97
triangle-tireworld-2014	6.18	4.25	5.00	4.80
wildfire-2014	9.02	9.67	9.47	9.69
academic-advising-2018	4.36	**7.42**	4.38	5.37
cooperative-recon-2018	**3.93**	1.52	2.25	0.67
Sum	71.12	73.79	69.23	72.71

Evaluation: Heuristics Performance

Table: Heuristic guidance

Domain	IDS	Propagation	Conformant
skill-teaching-2011	8.09	9.49	9.26
sysadmin-2011	5.11	9.21	9.24
tamarisk-2014	5.00	9.30	9.75
wildfire-2014	6.38	9.42	5.04
academic-advising-2018	0.77	4.49	3.32
Sum (all domains)	89.13	61.89	54.29

Table: Performed trials

Domain	IDS	Propagation	Conformant
sysadmin-2011	232'050	249'611	139'629
Sum (all domains)	1'490'326	2'948'572	1'649'386