Modified MCTS for Elevator Transportation

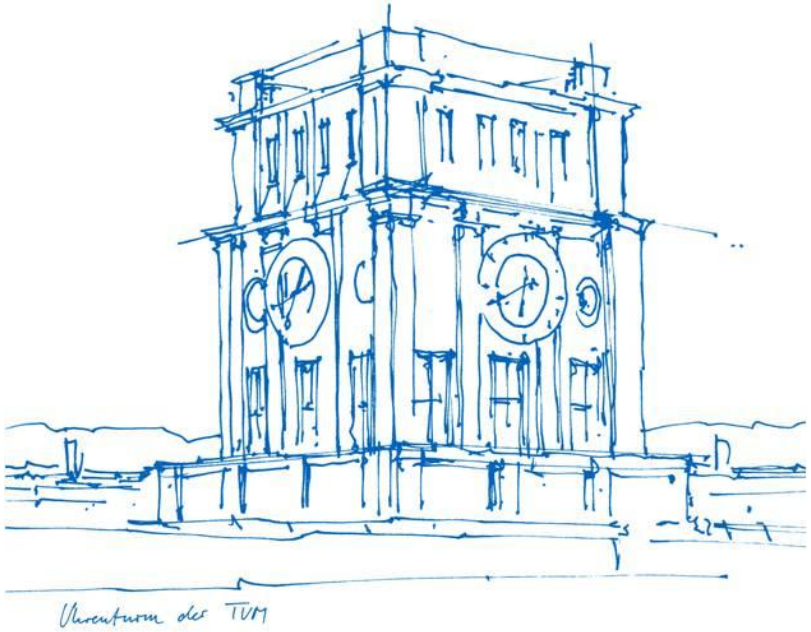
Group 5: Maximilian Rieger, Tim Pfeifle

Advisor: Sebastian Bachem

Technical University Munich

Advanced Deep-Learning in Robotics

Munich, 18. June 2020

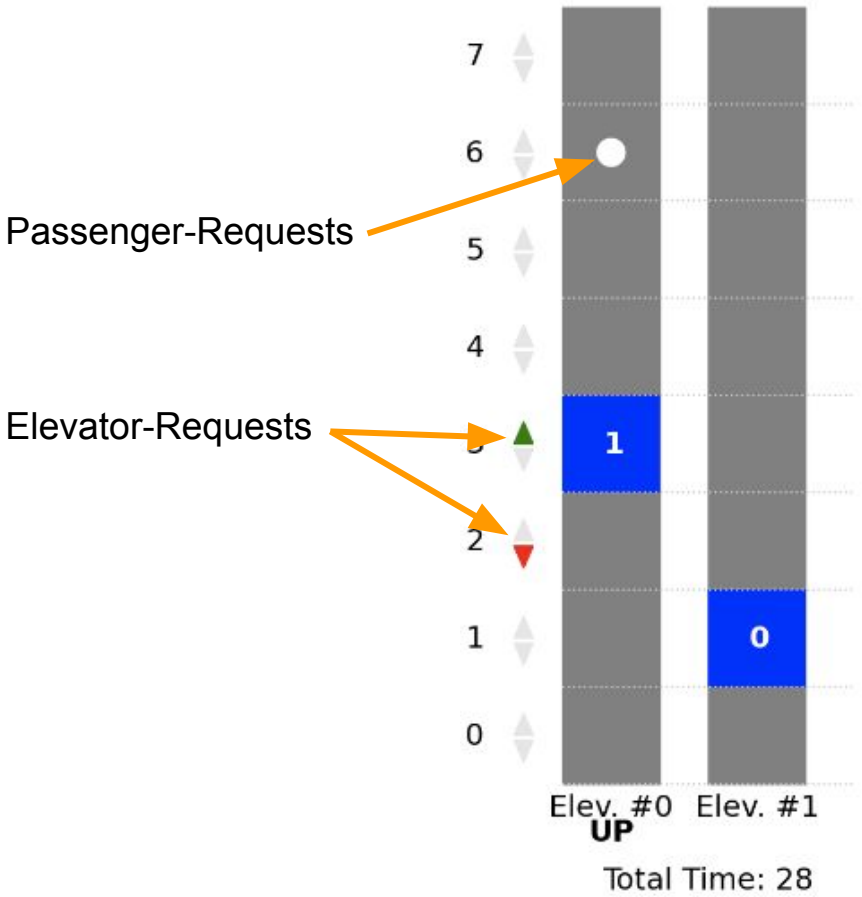


Elevator Transportation

3 Actions per Elevator:
- Up
- Down
- Open

Goal: Minimize Passenger Waiting Time

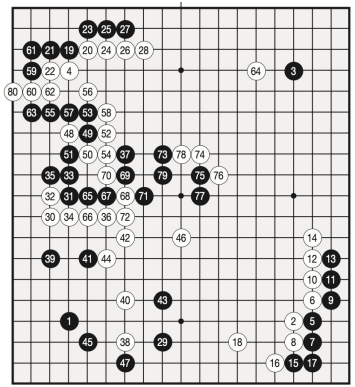
State not fully observable!



Why Elevator Transportation?

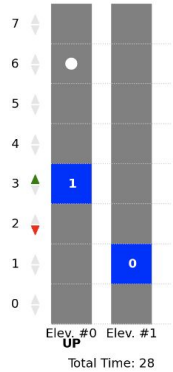
Modified AlphaZero Algorithm for new class of problems

Go



rewards at the end of the episode/game

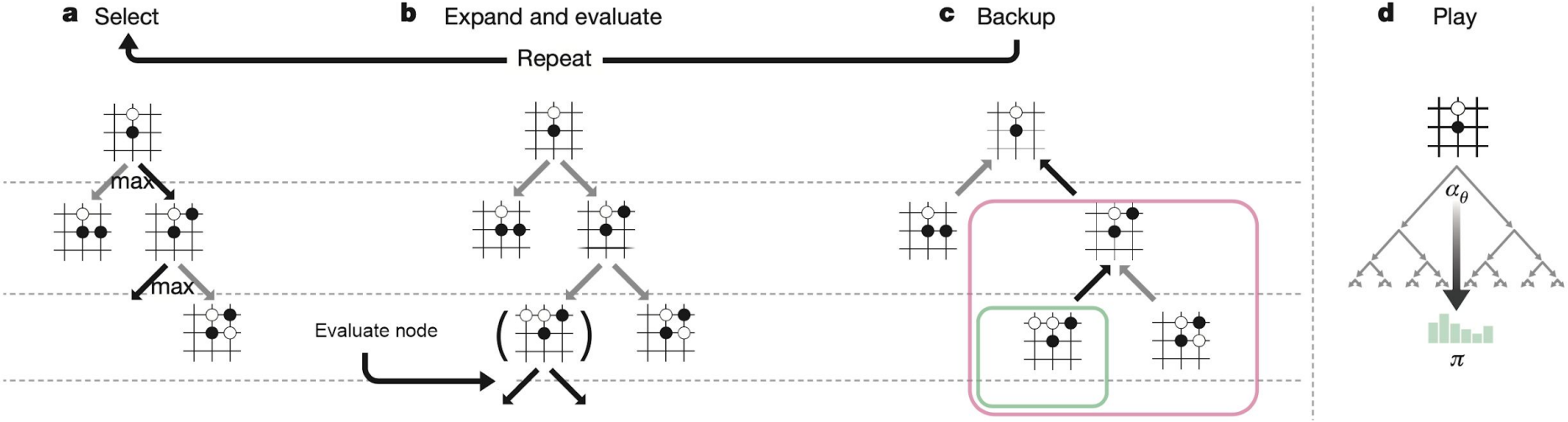
Elevator Transportation (Representative)



Passenger waiting-times

→ **observed rewards in every step**

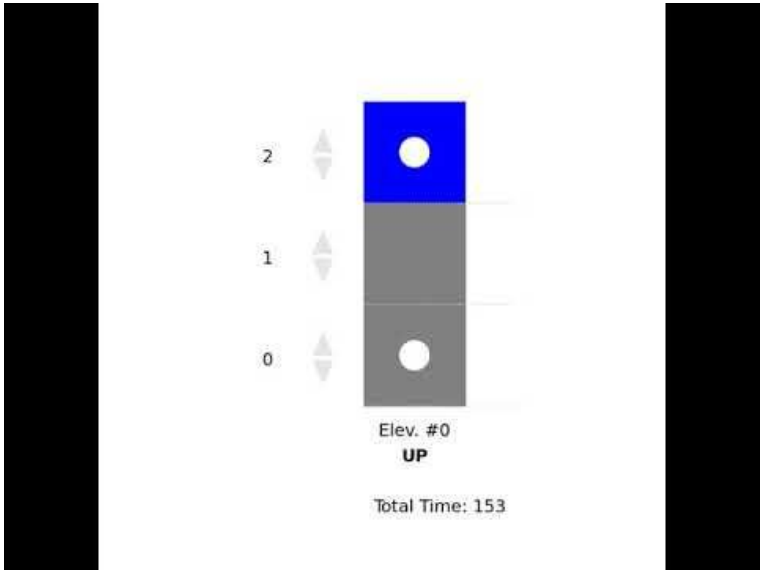
Monte-**C**arlo-**T**ree-**S**earch (AlphaZero [1])



Our Modification: Use observed rewards at each step

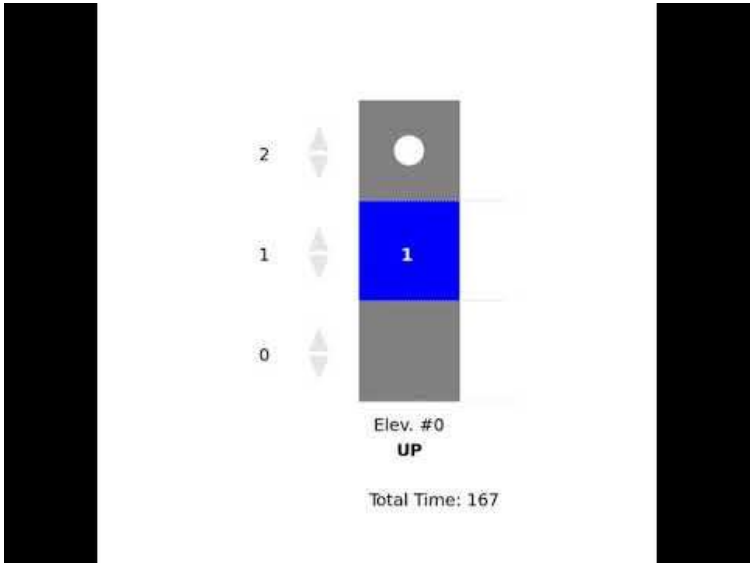
First results

Random Policy



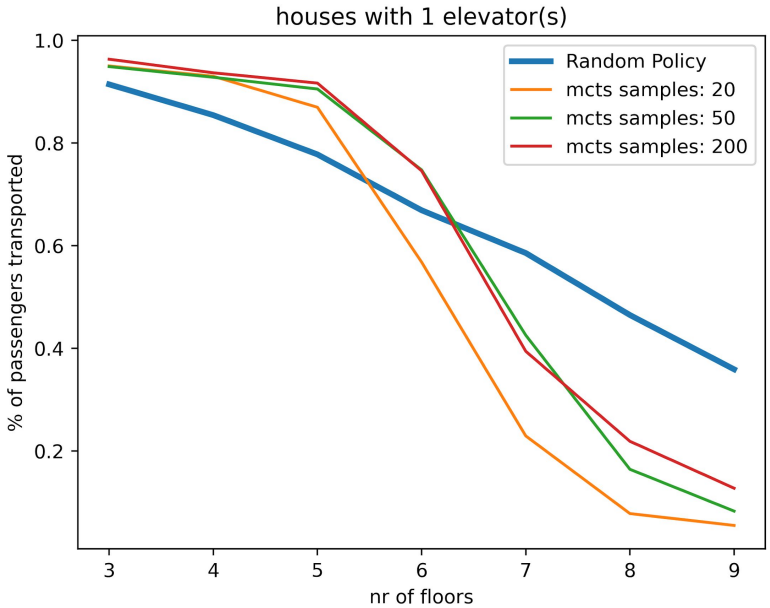
https://youtu.be/QkYXH6Dtejo

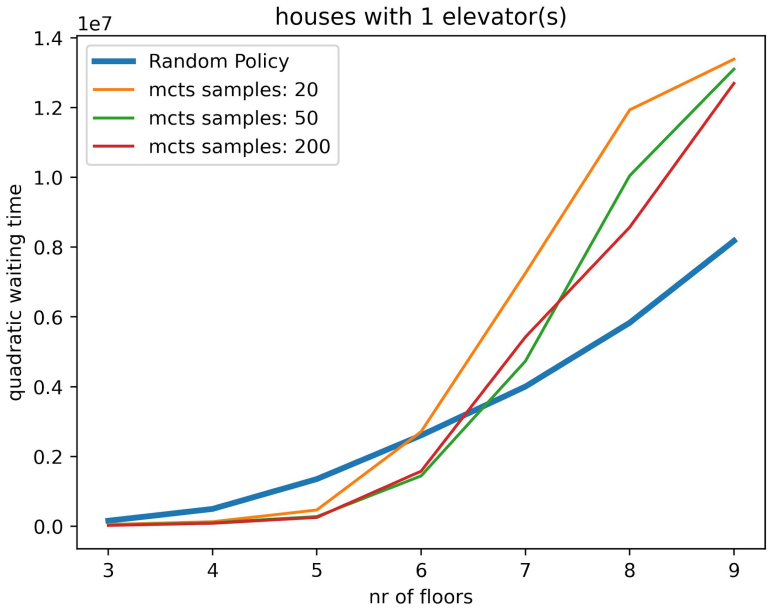
Modified-MCTS



https://youtu.be/gMstKN3pXR8

First results





Progress

Done so far

- Environment + Passenger-Generator ✓
- MCTS ✓
- Random Policy ✓
- Simple Model / Training Setup ✓

Next Steps

- Train Neural Network to guide MCTS (similar to AlphaZero)
- Compare to heuristic elevator baselines
- Ranked reward = $\begin{cases} +1 & \text{if result is better than 75\% of previous} \\ -1 & \text{else} \end{cases}$ [2]

Thank you!

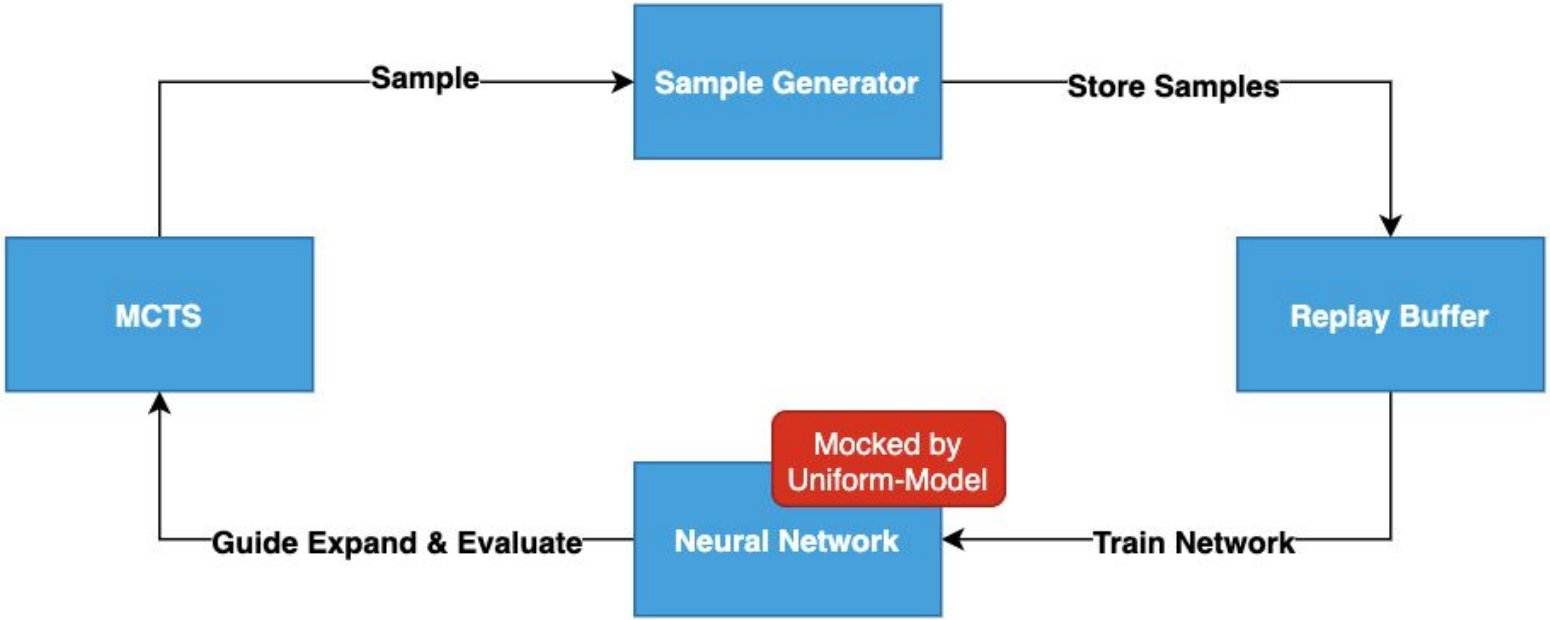
Q&A

References

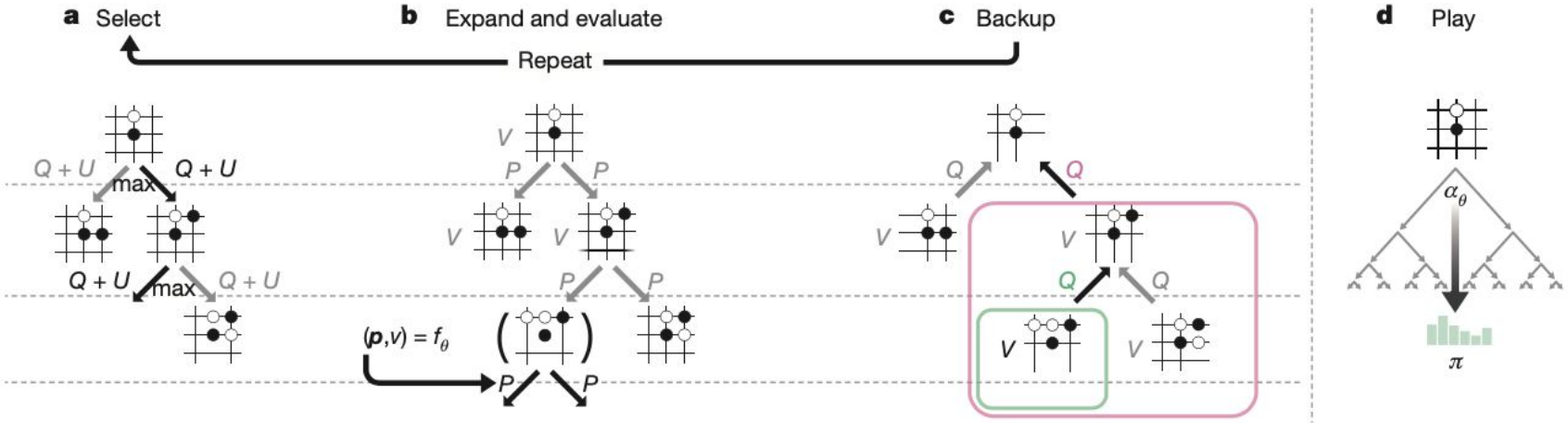
[1] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”. In: Science 362.6419 (2018), pp. 1140–1144

[2] A. Laterre, Y. Fu, M. K. Jabri, A.-S. Cohen, D. Kas, K. Hajjar, T. S. Dahl, A. Kerkeni, and K. Beguir. “Ranked reward: Enabling self-play reinforcement learning for combinatorial optimization”. In: arXiv preprint arXiv:1807.01672 (2018)

Algorithm (Overview)



MCTS (AlphaZero)



- Action Value: Q(s, a) How good is the action a?
- Upper Confidence Bound: U(s, a) Should I explore a further?
- Visit Counter: N(s, a) How often was a visited?

MCTS (AlphaZero) + Our Modification

$$a = \arg\max_a Q(s,a) + U(s,a)$$

$$Q(s,a) = \frac{1}{N(s,a)} \sum_{s'} v(s')$$

$$U(s,a) \propto \frac{\boldsymbol{p}(s,a)}{1 + N(s,a)}$$

$$Q_{new}(s,a) = \frac{1}{N(s,a)} \sum_{s'} c_{obs} \cdot f_{norm}\left(\frac{r(\pi_{s,s'})}{|\pi_{s,s'}|}\right) + (1 - c_{obs}) \cdot v(s')$$

Length of path from s to s'

$$f_{norm}(x) = \tanh\left(\frac{x}{10}\right)$$

- Action Value: Q(s, a) How good is the action a?
- Upper Confidence Bound: U(s, a) Should I explore a further?
- Visit Counter: N(s, a) How often was a visited?