Hand-in Discussions

Group 28

Rebecca Weiß
Maximilian Kleinegger
Maximilian Seeliger

Introduction

- Algorithm
- Chosen maps
- Explored paths
- Evaluation
- Drawbacks
- Discussion / Code

Algorithm - Monte Carlo

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-soft policy

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

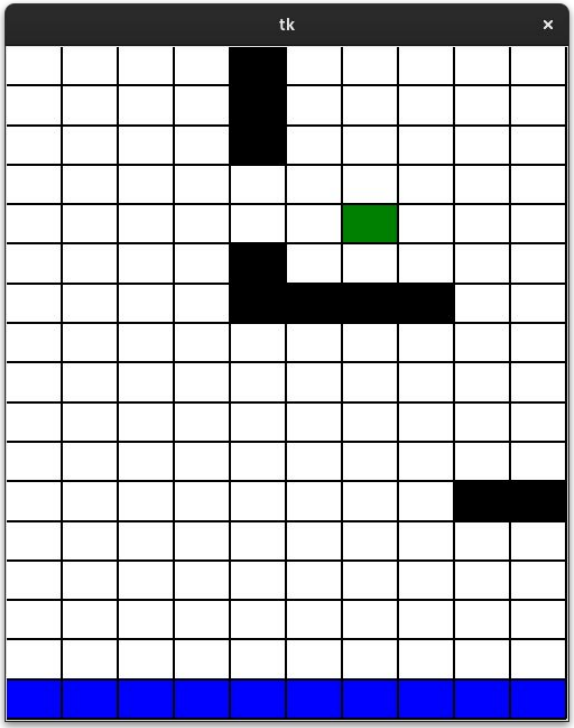
$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

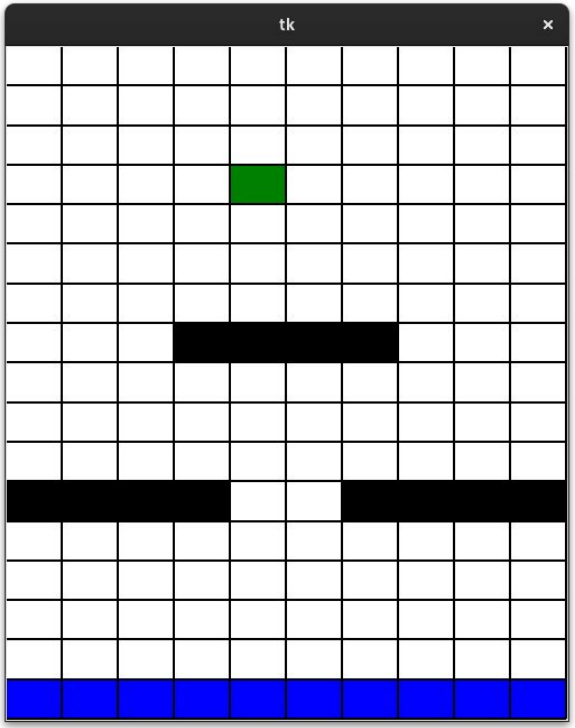
For all $a \in \mathcal{A}(S_t)$:

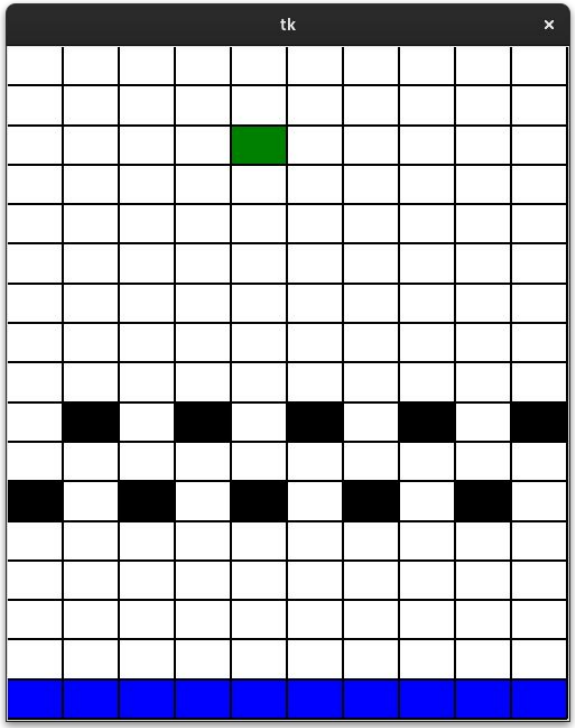
$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

Maps

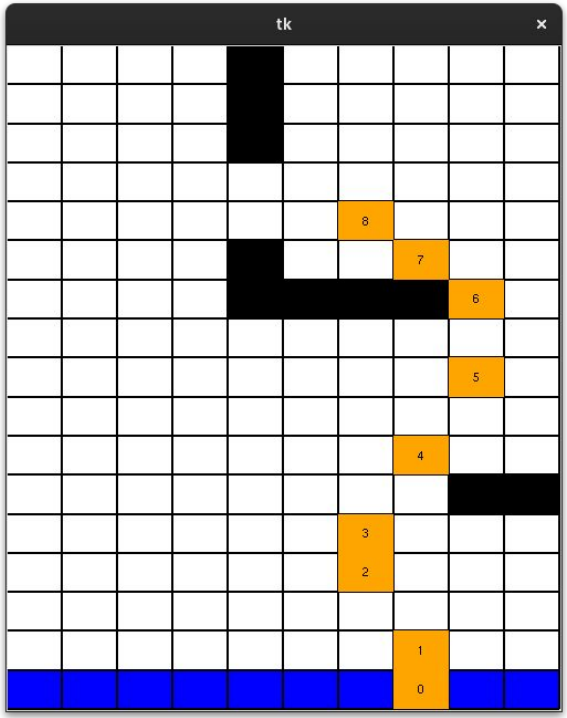
Also in
reduced form

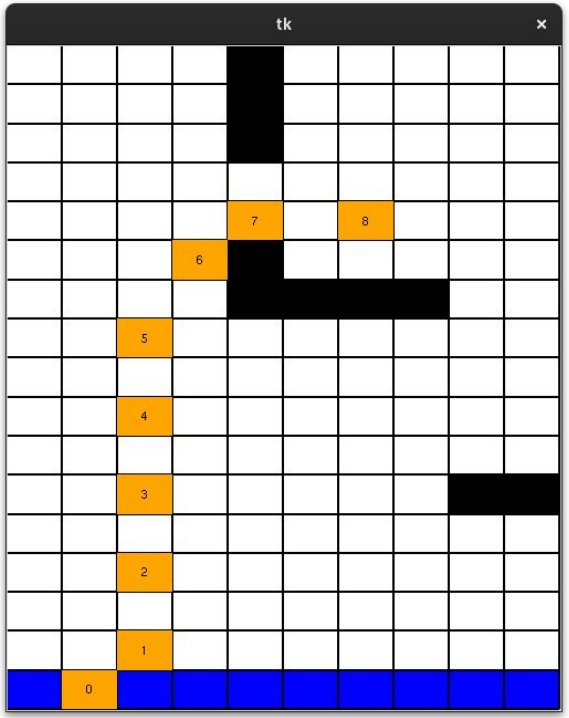




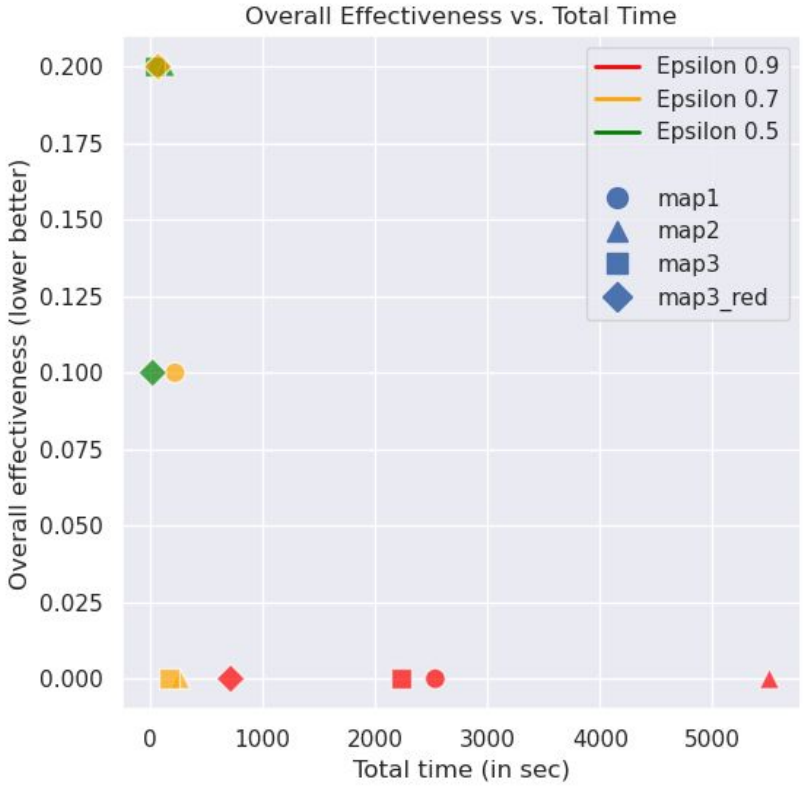


Examples of generated paths

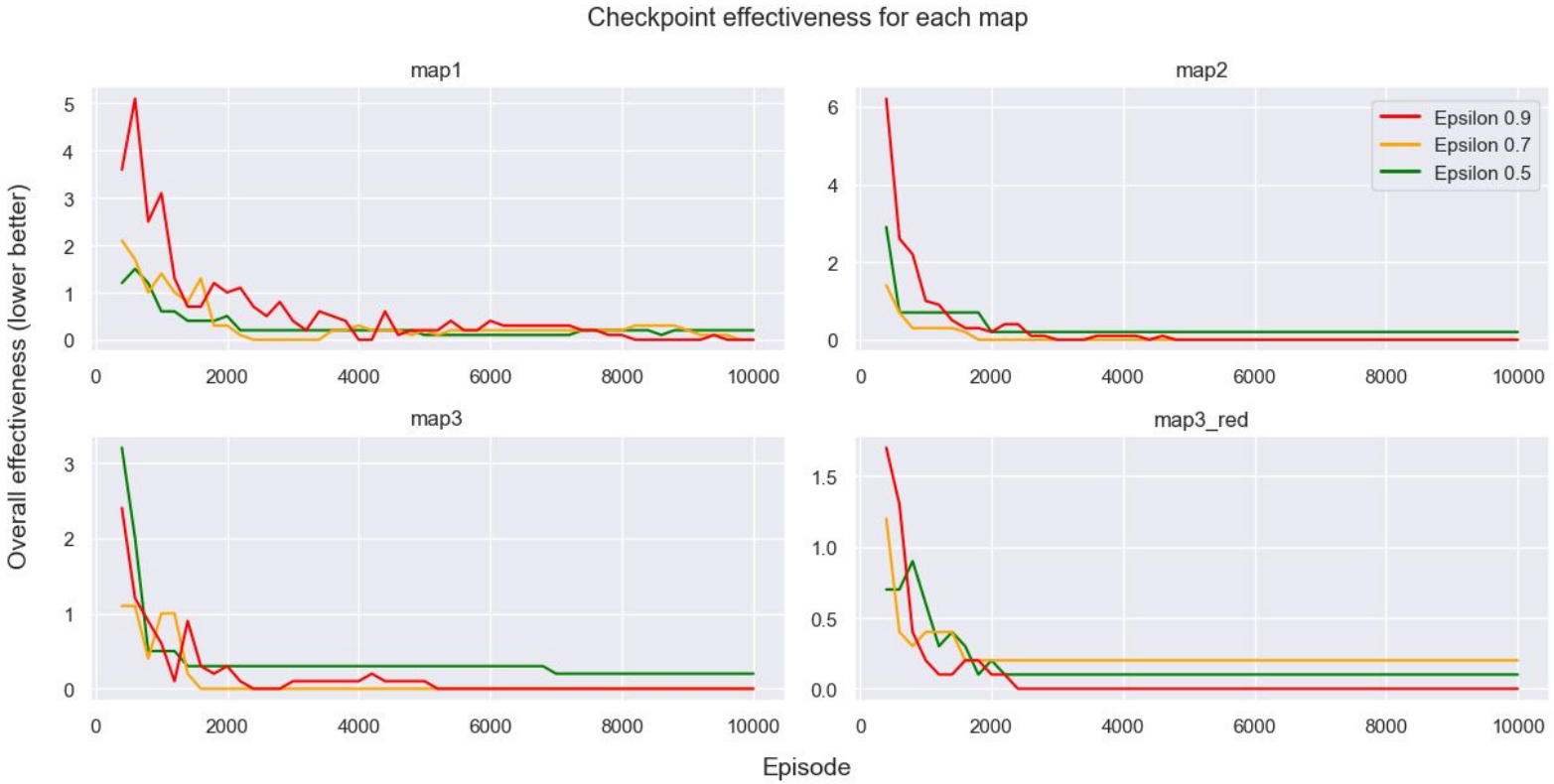




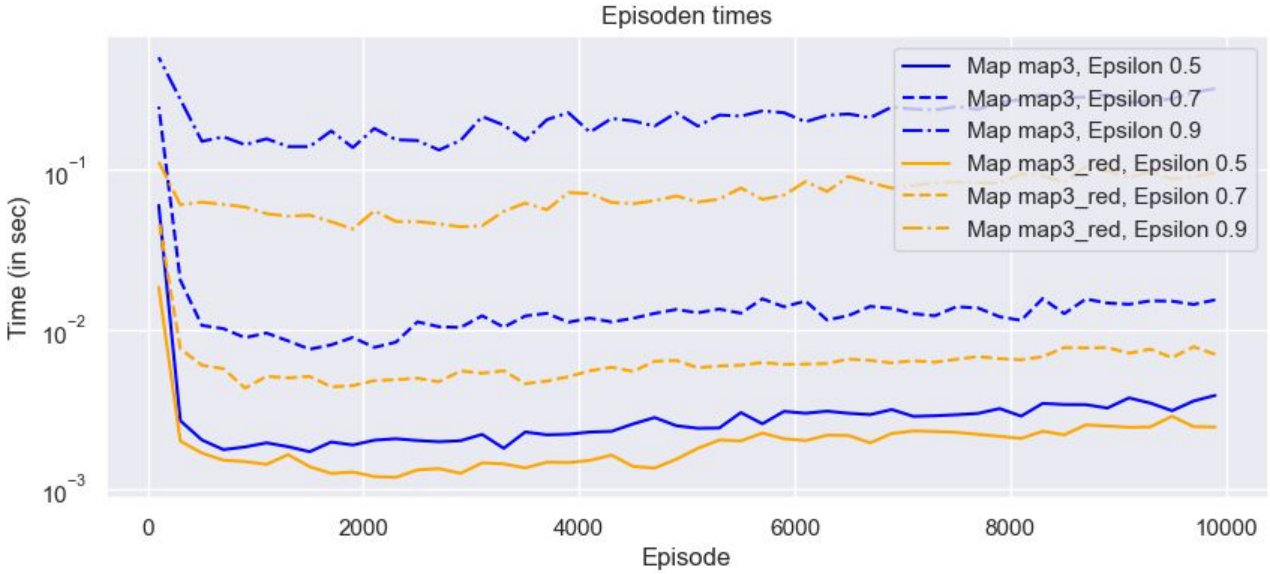
Effectiveness vs. Runtime



Effectiveness at pathfinding (MAE)



Runtime regarding #obstacles



Epsilon

- **Epsilon** is **fundamental** to the success and failure

- **High** epsilon -> focus on **exploration**
- **Low** epsilon -> focus on **exploitation**

- Agent **evaluation** only **after** training allows **high** epsilon

- **Training time decreases with epsilon**, as episodes are found more quickly based on previous knowledge

Drawbacks

- **No transductive** capabilities
- Episode definition with **random restarts** possibly leads to **unexplored starting positions** (agent moves to invalid position to be restarted at a better start)
- **Actions in second derivative** rather than first increases the problem complexity (large state-space, large action-space)
- **Evaluation** is very problem specific and task specific (quickly performing agent vs. specific training time)

Conclusion

- Exploration vs Exploitation
- Problem specific evaluation
- No transductive capabilities
- Complex problem definition