Exercise 5.10: ***Racetrack (programming)*** We have applied the off-policy Monte Carlo control method using weighted importance sampling to the racetrack problem. We have chosen $\varepsilon = 0.05$ for the behavior ε-soft policy b; the noise defined in the exercise was set to 0.1. The speed of the learning process was improved by updating the behavior policy b according to the optimal target policy π (so the actions to be selected by b with the largest probability would be identical to the actions given by the target policy π, see the code for details). For the settings, one million of episodes was not enough to find the optimal policy; therefore, we have used 100 million episodes for the training.

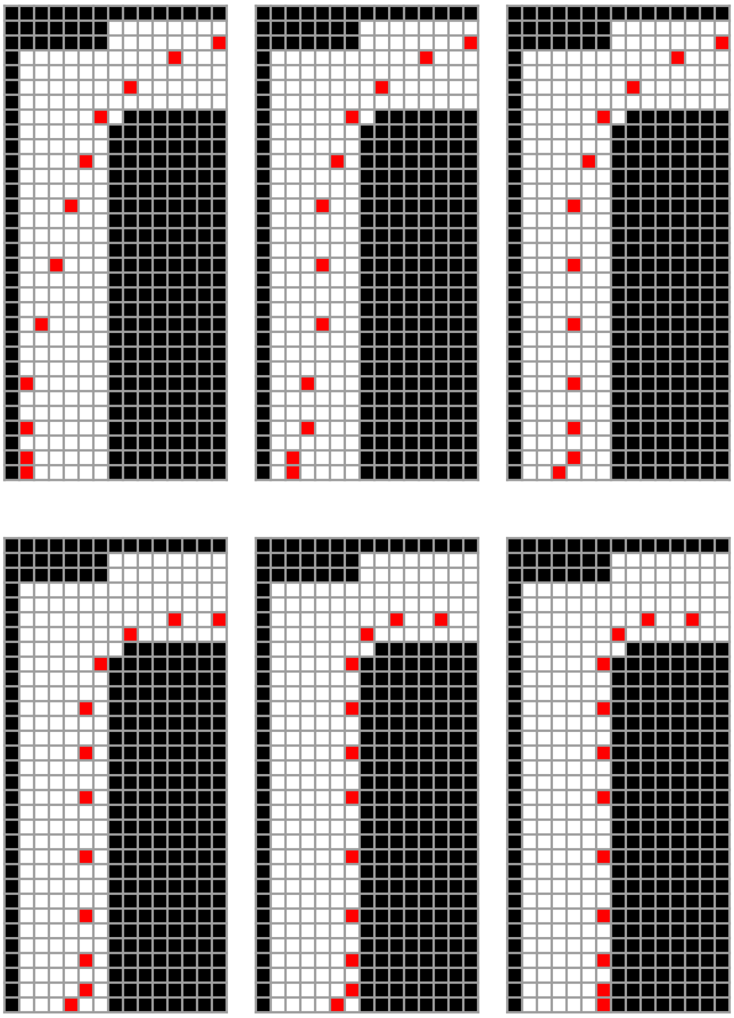

Figure 1: *Exercise 5.10*: Trajectories obtained by the off-policy Monte Carlo control method (based on the weighted importance sampling) trained on 100 million episodes.