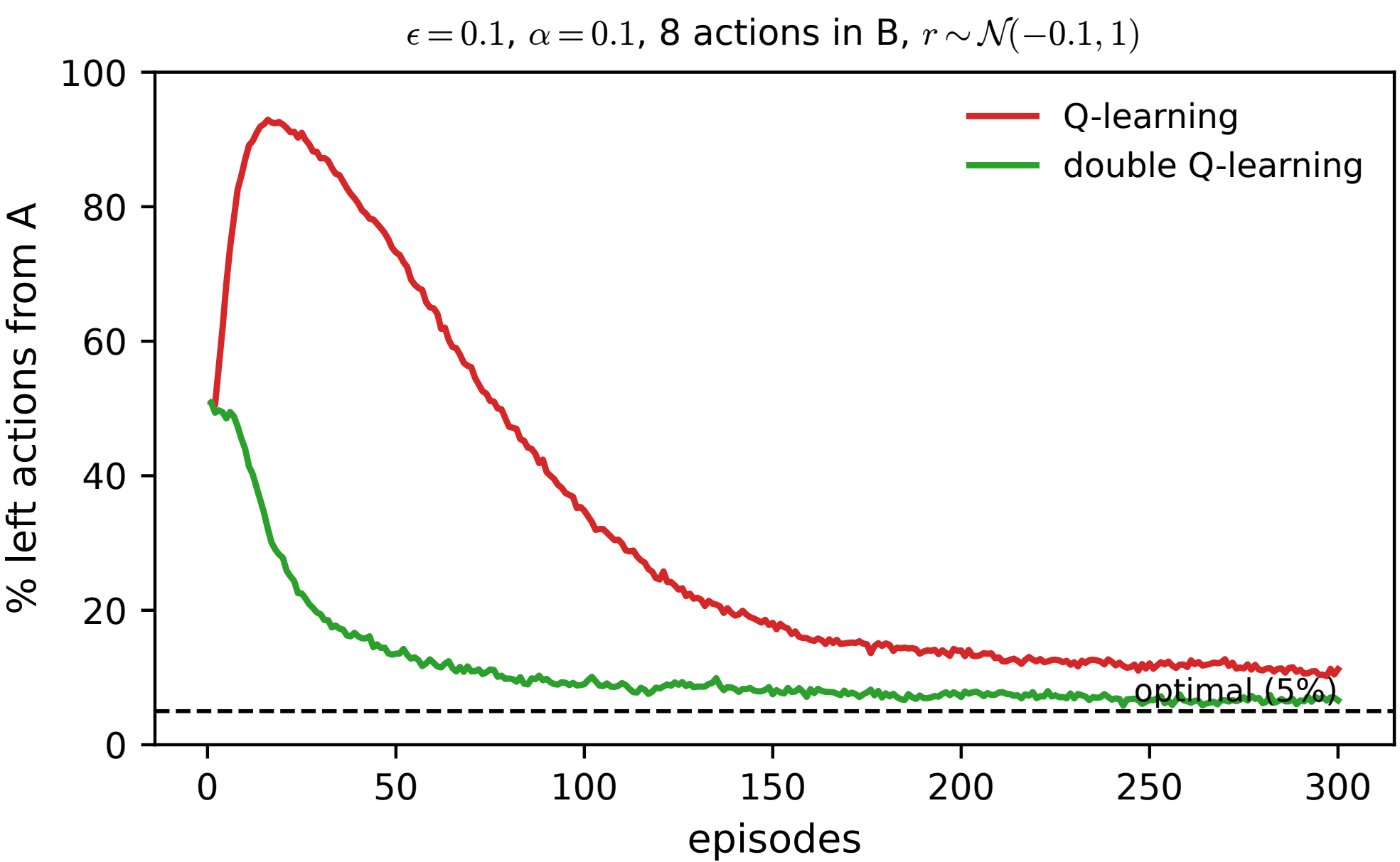

$\epsilon = 0.1$, $\alpha = 0.1$, 8 actions in B, $r \sim \mathcal{N}(-0.1, 1)$
Q-learning
double Q-learning
optimal (5%)
% left actions from A
episodes
0
20
40
60
80
100
0
50
100
150
200
250
300