Safe risk MDP Q-value, property R1, shallow 3-layered network 5 4 Average reward 3 One-hot encoding 1 Ordinal encoding One-hot encoding (Q-table) Ordinal encoding (Q-table) 0

300

Number of episodes

400

500

600

Modest solution = 4

200

100