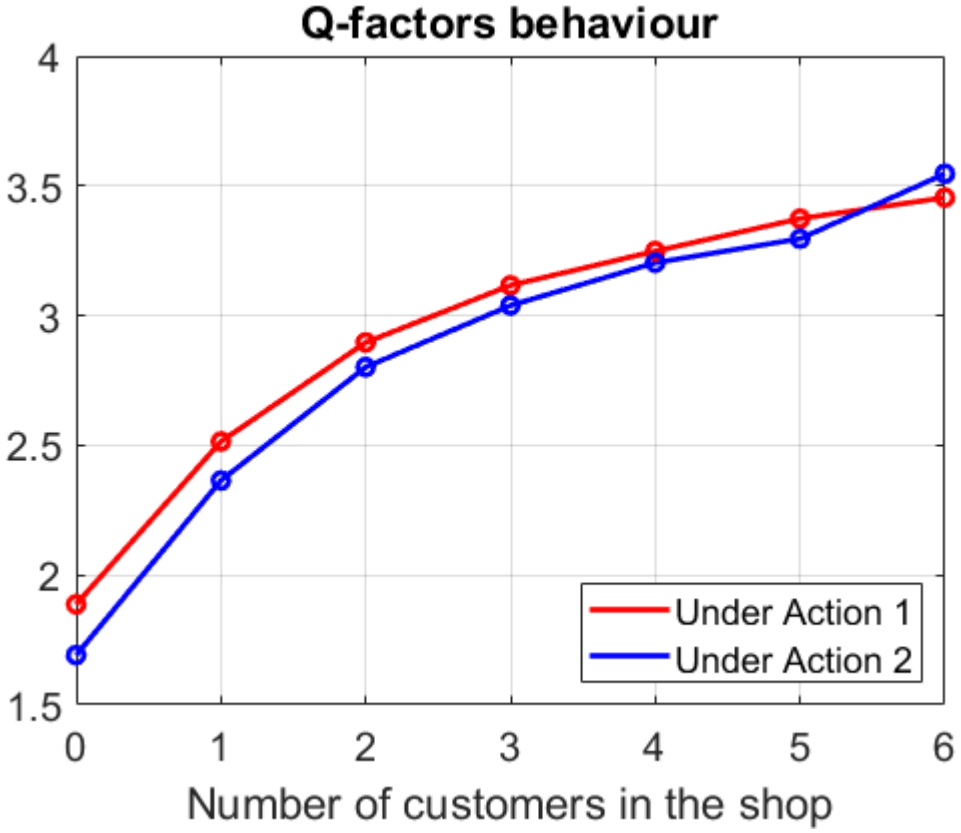
Part 2: Summary of the results.

Obviously results change at each trial, however they always appear to have a certain behaviour, explained by the resulting Q-factors:

$$Q = \begin{pmatrix} 1.8873 & 1.6914 \\ 2.5147 & 2.3643 \\ 2.8965 & 2.8014 \\ 3.1165 & 3.0389 \\ 3.2483 & 3.2038 \\ 3.3737 & 3.2963 \\ 3.4547 & 3.5464 \end{pmatrix}$$

Clearly, the vector of the best policy is $d = (1, 1, 1, 1, 1, 1, 2)$, which means that action 1 is the best choice if there are less than 6 customers, while action 2 is better if there are 6 customers.

The Q-factors here have the following behaviour:



It is obvious that the two curves are not decreasing. However, we notice that qualitatively the two "functions" get even more nearer and in the state 7 (6 customers in the shop) the Q-factor related to the action 2 is higher than the one related to action 1.