we see a similar outcome for average rewards. Optimistic and realistic greedy eventually converge, while UCB performs better due to ~~optimised~~ better exploration and ceasing exploration when a good estimate has been found.

Q4

2. In Non-stationary case.

Optimal Action: UCB performance ~~keeps~~ is bad because it stops exploring (≈2000) actions after ~~some~~ steps have passed.

It starts acting greedily and its estimates do not get updated with a constant step-size.

Optimistic initially performs well. This is due to optimal action being picked more often. However because of the initial bias, there will be significant variance in ~~the~~ estimates of actions other than the supposed "optimal" estimate if got early in the simulation.

Realistic eventually performs better because it keeps exploring and ~~the~~ samples the distributions is updated. throughout

Avg. Rewards : Optimistic performs best due to initial bias and selection of the optimal value.

UCB ~~performs better than realistic~~ because of ~~initially exploring~~