St+1

St

C\* C1 C2C3C4

C\*

C1

C2

C3

C4

For very short runs, the number of correctly pulled handles could be much lower than this value proposed the equation above. This is because the first few pulls assumes all handles have equal probability, until 1 handle has a higher estimated reward than all others, then it uses the ε-greedy strategy where the optimal handle will be pulled at much higher frequency and ‘explores’ the other at a lower frequency. However, because all the handles have variance in returned rewards, a sub-optimal handle may produce a higher reward than the optimal handle in short runs. As such, the sub-optimal handle will be pulled more often as the model assumes it is optimum. Since we select the best action to be over time, as exploration of other handles increases and the mean estimated reward averages out, the number of correctly pulled handles will converge on the long-term average equation.

From figure 1 below, it can be seen that initially that sub-optimal handles was pulled for a large number of time steps as it may have performed above average or optimal handle performed below average. However as other handles are explored over time, the variance will become less significant so the percentage of optimal pulls increases quickly until it converges on the long-term average probability value of .

Table

Description automatically generated

Chart, line chart

Description automatically generated

Figure 1 shows the percentage of optimal pulls as the equation begins to converge for 10,000 and 100,000 pulls respectively