The bandit problem is extended by implementing Monte Carlo analysis. In this method, we run the whole exercise multiple times and average every time step. We then select the action that produces the best reward from this average and exploits it from then onwards.

Figure 1 shows the cumulative regret after running such algorithm. As can be seen, fewer runs results in greater cumulative regret, and for lower values (1,10) the regret increases linearly. This is because the rewards from each station, has a gaussian distribution so it may be for instance a suboptimal charging station returned a higher reward for the short amount of trials it had.

On the other hand for higher number of runs for example 2000, the regret increases linearly at first, before suddenly having an almost constant regret. This is because Monte Carlo analysis takes into account the training regret as well as the regret that occurs after training. Therefore at first, for 2000 runs, all the charging stations are tried. Once the number of runs have been attempted, it exploits the what it believes to be optimal which in this case it is.

The regret in this case is not fixed because each charging station has quite a high variance return in comparison to the mean so the regret will continuously fluctuate.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 1: Plot of cumulative regret for different number of runs

By varying the values of epsilon and number of Monte Carlo runs, the long term average probability equation can be proven to be correct as can been seen in figure 3, the percentage of optimal pulls converges on the long-term average probability value of shown in figure 2.

The larger values of epsilon converge on a much lower percentage optimal actions pulled. This is because large values of epsilon result in more exploration, most clearly explained when epsilon = 1, which means there is no exploitation at all, only exploration and since there are 4 different stations to choose from, the random long term probability of picking each station is .

Text

Description automatically generatedTable

Description automatically generated

Figure 2: Graph of percentage of optimal pulls using e-greedy strategy and its long term converged value for each value of epsilon

Chart, line chart

Description automatically generatedThis can also be justified through the cumulative regret which shows that the lower the value of epsilon, the lower the cumulative regret over time. Furthermore, the regret increases linearly over time which confirms that the exploration term never ends for all values of epsilon greater than zero and the system will always explore with a probability of epsilon (figure 3).

Figure 3: Graph of percentage cumulative regret for different values of epsilon