The issue with e-greedy is its constant exploration term, even in the long term, so even if we develop a strong understanding of the environment, it will never only exploit the optimal value. As a result, the regret increases linearly for all time steps. Therefore epsilon greedy can be refined using various strategies to improve its performance.

A method of doing this is tempering exploration. What this means is that we make the exploration probability ε, a non-increasing function of time εt and is defined mathematically as: . So over time, the probability of exploring decreases so we exploit the optimal solution more. Action distribution is shown below.

An example of this is a damped ε-greedy strategy shown in equation X, where c is a confidence value chosen by the user to determine how long the system should explore for. As the number of runs increases, the value of ε decreases eventually becoming 0 meaning no exploration.

(X)

The performance of this algorithm can be shown below (figure 1), the graph below shows the percentage of optimal actions taken. We can see that clearly, the long term average converges on the optimal action.

Chart

Description automatically generated with low confidence

Figure 1: Plot of percentage of optimal actions taken

However, with this strategy, faulty cases often occur due the distribution of rewards for each charging station, a suboptimal solution may be picked as it performed better than the optimal solution and since over time, the exploration decreases, the optimal solution may be completely missed and the system results in only exploiting a suboptimal solution (figure 2).

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 2: Plot of percentage of optimal actions taken with faulty cases

Finally, we can see the advantages and disadvantages of using such an algorithm with the graph below showing the cumulative regret (figure 3). If an action is selected correctly, the regret is almost none as it only exploits the optimal action. However if a suboptimal action is picked, the regret increases linearly as it only exploits a suboptimal action.

Chart, line chart

Description automatically generated

Figure 3: Plot of cumulative regret