Arshia Zakeri Rad
Tigran Tumanyan

Report 1: UCB Algorithm - Experimental Observations

Observations:

The UCB algorithm was evaluated over 100 independent environments, each with 5000 interaction steps. In all worlds, the algorithm initially explores different actions, which results in a relatively low average reward and a small number of optimal action selections during the early iterations.

As time progresses, the number of optimal action selections increases steadily, and the average reward consistently improves. For example, in later iterations (around 4000–5000 steps), many environments achieve an average reward close to or above 0.9, indicating that the algorithm has successfully identified and exploited the optimal action. In World 99, the number of optimal action selections reached 3932 by iteration 5000, with an average reward of approximately 0.91.

However, the learning speed varies across environments. In some worlds, such as World 100, the algorithm converges more slowly and achieves a lower final average reward (around 0.69). This behavior occurs when the difference between the optimal action and suboptimal actions is small, making it harder for the algorithm to distinguish between them.

Overall, the results demonstrate that UCB reliably improves performance over time and converges toward selecting the optimal action, though the convergence rate depends on the difficulty of the environment.