2.1 In the comparison shown in Figure 2.1, which method will perform best in the long run in terms of cumulative reward and cumulative probability of selecting the best action? How much better will it be?

SOLUTION According to the figure, the e-greedy method with epsilon=0.1 performs best in terms of cumulative reward and cumulative probability of selecting the best action in the long run. Although the e-greedy method with epsilon=0.01 takes longer to converge, it eventually outperforms the other methods in both measures.

The greedy method performs the worst in the long run, achieving a reward per step of only about 1 compared with the best possible of about 1.55 on this testbed. It also selects the optimal action in only approximately one-third of the tasks.

In contrast, the e-greedy methods eventually perform better because they continue to explore, improving their chances of recognizing the optimal action. The e-greedy method with epsilon=0.1 explores more and usually finds the optimal action earlier, but never selects it more than 91% of the time. The e-greedy method with epsilon=0.01 improves more slowly but eventually performs better than the e-greedy method with epsilon=0.1 on both measures.

In terms of the magnitude of improvement, the difference between the worst-performing method (greedy) and the best- performing method (e-greedy with epsilon=0.01) is significant. The e-greedy method with epsilon=0.01 achieves a reward per step of approximately 1.4, which is almost 40% better than the greedy method's performance. Similarly, the e-greedy method with epsilon=0.01 selects the optimal action in approximately 70% of the tasks, which is more than twice as often as the greedy method's performance.