1. (Convergence of temporal difference value learning) In TD value learning, we repeatedly try out a given policy. At the nth trial, we collect a sampled value x_n and use it to update our value estimate V_n. Suppose we use a time-varying learning rate of $\alpha_n = 1/n^2$ to compute the exponential moving average, and assume that both $\{x_n\}$ and $\{V_n\}$ are bounded sequences, i.e, $\forall n > 0, |x_n| \leq C_1$ and $|V_n| \leq C_2$ for some constants $C_1 > 0$ and $C_2 > 0$. Prove that TD value learning using this particular choice of learning rates $\{\alpha_n\}$ will converge. Hint: Verify that the sequence $\{V_n\}$ with $V_n = (1 - \alpha_n) V_{n-1} + \alpha_n x_n$ is a Cauchy sequence.

2. (Implementation of the SARSA and Q-learning algorithms) You are required to implement the SARSA and Q-learning algorithms for the cliff-walking problem. In this problem, the agent starts in the bottom left corner and is expected to reach the bottom right corner. Stepping into the cliff that segregates those tiles yields a massive negative reward and ends the episode. Otherwise, each step comes at a small cost, meaning the shortest path is the optimal policy (please see codes for more details). You are required

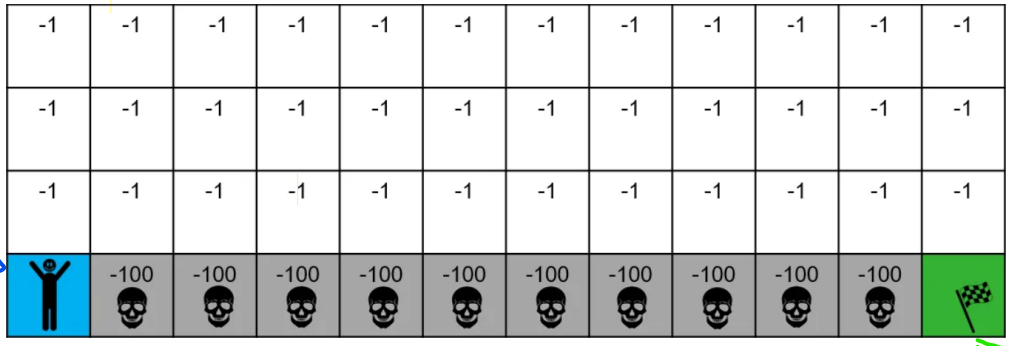

Figure 1: The cliff walking example.

to first implement the ϵ-greedy policy in `take_action()` function of `base.py`. Then implement the `update()` function, which updates the Q-table of Q-learning, and the `cliff_walk()` function, which instantiates the Q-learning algorithm, in `QLearning.py`. At last, implement the `update()` function, which updates the Q-table of SARSA, and the `cliff_walk()` function, which instantiates the SARSA algorithm, in `SARSA.py`. Run your codes to show

   (a) the performance of SARSA algorithm with different values of ϵ in ϵ-greedy;

   (b) the performance of Q-learning algorithm with different values of ϵ in ϵ-greedy;

   (c) the performance of Q-learning algorithm with different values of ϵ in ϵ-greedy but using the **target policy** (*i.e.*, the learned optimal policy in each episode).

   Compare the performance of the algorithms in the above cases and give discussions about

   (a) what are the impacts of different values of ϵ on the performance of the above three algorithms?

   (b) what is the difference between the performance of the behavior policy of Q-learning algorithm and the performance of the target policy of Q-learning algorithm?