Implementation: Sarsa(0)

The pseudocode for Sarsa (or Sarsa(0)) can be found below.

TD Control: Sarsa(0)

Input: policy π, positive integer *num_episodes*, small positive fraction α
Output: value function Q ($\approx q_\pi$ if *num_episodes* is large enough)
Initialize Q arbitrarily (e.g., $Q(s,a) = 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$, and $Q(\textit{terminal-state}, \cdot) = 0$)
for $i \leftarrow 1$ ***to*** *num_episodes* **do**
    $\epsilon \leftarrow \frac{1}{i}$
    Observe S_0
    Choose action A_0 using policy derived from Q (e.g., ϵ-greedy)
    $t \leftarrow 0$
    repeat
        Take action A_t and observe R_{t+1}, S_{t+1}
        Choose action A_{t+1} using policy derived from Q (e.g., ϵ-greedy)
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$
        $t \leftarrow t + 1$
    until S_t *is terminal*;
end
return Q

Sarsa(0) is **guaranteed to converge** to the optimal action-value function, as long as the step-size parameter α is sufficiently small, and the **Greedy in the Limit with Infinite Exploration (GLIE)** conditions are met. The GLIE conditions were introduced in the previous lesson, when we learned about MC control. Although there are many ways to satisfy the GLIE conditions, one method involves gradually decaying the value of ϵ when constructing ϵ-greedy policies.

In particular, let ϵ_i correspond to the i-th time step. Then, if we set ϵ_i such that:

- $\epsilon_i > 0$ for all time steps i, and
- ϵ_i decays to zero in the limit as the time step i approaches infinity (that is, $\lim_{i\to\infty} \epsilon_i = 0$),

then the algorithm is guaranteed to yield a good estimate for q_*, as long as we run the algorithm for long enough. A corresponding optimal policy π_* can then be quickly

Please use the next concept to complete **Part 2: TD Control: Sarsa** of `Temporal_Difference.ipynb`. Remember to save your work!

If you'd like to reference the pseudocode while working on the notebook, you are encouraged to open this sheet in a new window.

Feel free to check your solution by looking at the corresponding section in `Temporal_Difference_Solution.ipynb`.

NEXT