Deep Q Learning

Just like Deep SARSA, we have a neural network to predict the Q value of state-actions given a state input. But we will have 2 seperates policies: b which is an ϵ-greedy policy for exploration and π which is a greedy (target) policy for exploitation. The target policy will sample from the target network when estimating the Q values for bootstrapping:

$$\hat{L}(w) = \frac{1}{|K|} \sum_{i=1}^{|K|} [R_i + \gamma \max_a \hat{q}(S_i', A_i' | \theta_{target}) - \hat{q}(S_i, A_i | \theta)]^2. \quad (1)$$

where $|K|$ is the training batch size, and $\pi(s|\theta_{target}) = \max_a \hat{q}(S_i', A_i' | \theta_{target})$ is the action chosen by the target policy from the target network.

Afterwards, all the other steps regarding Replay Buffer sampling, Stochastic Gradient Descent, and cloning and updating target network is the same with Deep SARSA.