The 2 graphs displayed was from my original algorithm which is after each episode, update for all existed q(s,a).

Then I realized that I could have a faster algorithm: I can only update the policies for these states which appeared in the episode. Then, I modified my algorithm.

I think both methods make sense to me, however the graph of 2nd algorithm will squeeze more for episodes at the beginning. (In this case, it's the line for episode 100)