Sarsa: $R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$ — A_{t+1} selected on-policy (e.g. ε-greedy)

target

π must be b — action taken

Expected Sarsa: $R_{t+1} + \gamma \underbrace{E_\pi [Q(S_{t+1}, A')]}$

target

$$\sum_{a' \in A} \pi(a' | S_{t+1}) Q(S_{t+1}, a')$$

π does not need to be b.

↖ does not include A_{t+1}

Q-learning: $R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a')$

↑ does not include A_{t+1}