

## Off-policy $n$ -step Sarsa for estimating $Q \approx q_*$ or $q_\pi$

Input: an arbitrary behavior policy  $b$  such that  $b(a|s) > 0$ , for all  $s \in \mathcal{S}, a \in \mathcal{A}$

Initialize  $Q(s, a)$  arbitrarily, for all  $s \in \mathcal{S}, a \in \mathcal{A}$

Initialize  $\pi$  to be greedy with respect to  $Q$ , or as a fixed given policy

Algorithm parameters: step size  $\alpha \in (0, 1]$ , a positive integer  $n$

All store and access operations (for  $S_t$ ,  $A_t$ , and  $R_t$ ) can take their index mod  $n + 1$

Loop for each episode:

    Initialize and store  $S_0 \neq$  terminal

    Select and store an action  $A_0 \sim b(\cdot|S_0)$

$T \leftarrow \infty$

    Loop for  $t = 0, 1, 2, \dots$ :

        If  $t < T$ , then:

            Take action  $A_t$

            Observe and store the next reward as  $R_{t+1}$  and the next state as  $S_{t+1}$

            If  $S_{t+1}$  is terminal, then:

$T \leftarrow t + 1$

            else:

                Select and store an action  $A_{t+1} \sim b(\cdot|S_{t+1})$

$\tau \leftarrow t - n + 1$  ( $\tau$  is the time whose estimate is being updated)

                If  $\tau \geq 0$ :

$\rho \leftarrow \prod_{i=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_i|S_i)}{b(A_i|S_i)} \quad (\rho_{\tau+1:t+n-1})$

$G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$

                    If  $\tau + n < T$ , then:  $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n}) \quad (G_{\tau:\tau+n})$

$Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha \rho [G - Q(S_\tau, A_\tau)]$

                    If  $\pi$  is being learned, then ensure that  $\pi(\cdot|S_\tau)$  is greedy wrt  $Q$

        Until  $\tau = T - 1$