Algorithm parameter: small $\varepsilon > 0$ Initialize:

On-policy first-visit MC control (for ε -soft policies), estimates $\pi \approx \pi_*$

 $\pi \leftarrow \text{an arbitrary } \varepsilon\text{-soft policy}$ $Q(s, a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$$Returns(s, a) \leftarrow \text{empty list, for all } s \in \mathcal{S}, \ a \in \mathcal{A}(s)$$

Repeat forever (for each episode):

Repeat forever (for each episode):
Generate an episode following
$$\pi: S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$$

 $G \leftarrow 0$

$$G \leftarrow 0$$

Loop for each step of episode, $t = T - 1, T - 2, \dots, 0$:

 $G \leftarrow \gamma G + R_{t+1}$

For all $a \in \mathcal{A}(S_t)$:

op for each step of episode,
$$t = 1 - 1, 1 - 2, ...,$$

 $G \leftarrow \gamma G + R_{t+1}$
Append G to $Returns(S_t, A_t)$

Append G to $Returns(S_t, A_t)$

to
$$Returns(S_t, A_t)$$

 $\leftarrow average(Returns(S_t, A_t))$

 $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

If G to
$$Returns(S_t, A_t)$$

 $A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

 $A^* \leftarrow \operatorname{argmax}_a Q(S_t, a)$

$$ns(S_t, A_t)$$
 (with ties broken arbitrarily

$$(S_t, A_t)$$
 (with ties broken arbitrarily)

$$(\text{with ties broken arbitraril})$$

 $\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$