Off-policy MC control, for estimating $\pi \approx \pi_*$ Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:

 $Q(s,a) \in \mathbb{R}$ (arbitrarily)

 $C(s,a) \leftarrow 0$

 $\pi(s) \leftarrow \operatorname{arg\,max}_a Q(s, a)$ (with ties broken consistently)

Loop forever (for each episode):
$$b \leftarrow$$
 any soft policy

 $W \leftarrow W \frac{1}{b(A_{+}|S_{+})}$

Generate an episode using b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

episode using
$$b$$
:

 $G \leftarrow 0$ $W \leftarrow 1$

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

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

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

$$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$$

$$O(S, A_{\bullet})$$

$$\sum_{t} [G - Q(S_t, A_t)]$$

$$[C \quad O(S \quad A)]$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$$

$$f[G-Q(S_t,A_t)]$$

$$\frac{1}{2} [G - Q(S_t, A_t)]$$

th ties broken consistently)

$$(S_t, A_t) = (G - Q(S_t, A_t))$$

 $(S_t, A_t) = (With ties broken constant)$

$$\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$$
 (with ties broken consistently)

$$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$$
 (with ties broken consistently)
If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)