# Ch5 Monte Carlo methods

Section 5.1 to 5.4

- for estimating / learning the 'value function' and finding the optimal policy
- doesn't need prior knowledge, just experience

## 5.1 Monte Carlo Prediction

first visit — every visit

$v_\pi(s)$ average returns following first visit to s — avg. ret.

Backup diagram

**First-visit MC prediction, for estimating $V \approx v_\pi$**

Input: a policy $\pi$ to be evaluated
Initialize:
    $V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$ (Random value function)
    $Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):
    Generate an episode following $\pi$: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        Unless $S_t$ appears in $S_0, S_1, \ldots, S_{t-1}$: (ignore first-visit, remove line for every visit)
            Append $G$ to $Returns(S_t)$
            $V(S_t) \leftarrow$ average$(Returns(S_t))$

## 5.2 Monte Carlo Estimation for action values

need to estimate $q^*$

$q_\pi(s, a)$

first visit — every visit

converge quadratically

general problem: Maintaining Exploration

solve

"exploring starts"
every (s,a) pair has a non zero probability of being selected as the start

consider only stochastic policies w/ non zero prob for all possible actions for every state

## 5.3 Monte Carlo Control

- something like GPI (Generalized policy iteration)
- consider $\pi_0 \xrightarrow{E} q_{\pi_0} \xrightarrow{I} \pi_1 \rightarrow \dots \rightarrow \pi_* \xrightarrow{E} q_{\pi_*}$

complete policy evaluation
many episodes experienced

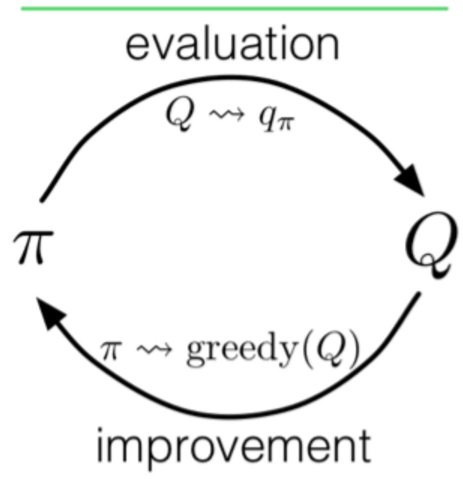
for each $s \in S$

$$\pi(s) = \underset{a}{\operatorname{argmax}}\, q(s,a)$$

Policy improvement theorem

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \underset{a}{\operatorname{argmax}}\, q_{\pi_k}(s,a))$$
$$= \max_a q_{\pi_k}(s,a)$$
$$\geq q_{\pi_k}(s, \pi_k(s))$$
$$\geq v_{\pi_k}(s)$$

- assumptions → exploring starts
  → $\infty$ steps policy eval




solve

consider some error, many episodes, approximate $q_{\pi_k}(s,a)$

- give up on finding $q_{\pi_k}(s,a)$ before I
  just "move toward it" like in value iteration
- in MC, natural to alternate b/w E and I on an episode by episode basis

**Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$**

Initialize:

Initialize:
$\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$ random policy
$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability $> 0$
Generate an episode from $S_0, A_0$, following $\pi$: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
$G \leftarrow \gamma G + R_{t+1}$
Unless the pair $S_t, A_t$ appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$: (first visit)
Append $G$ to $Returns(S_t, A_t)$
$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (policy improvement)

→ 5.4 Monte Carlo Control w/o Exploring Starts

off-policy

on policy → gen soft policy
eg MC control method

$\pi(a/s) > 0 \quad \forall s \in S$
$\forall a \in A(s)$

shifted to a deterministic one

$\varepsilon$ soft policy $\quad \pi(a/s) \geq \frac{\varepsilon}{|A(s)|}$

(a type
$\varepsilon$ greedy policy

$$\pi\left(\frac{a}{s}\right) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|A(s)|} & a^* \\ \frac{\varepsilon}{|A(s)|} & \text{else} \end{cases}$$

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

Algorithm parameter: small $\varepsilon > 0$
Initialize:
$\pi \leftarrow$ an arbitrary $\varepsilon$-soft policy
$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):
Generate an episode following $\pi$: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
$G \leftarrow \gamma G + R_{t+1}$
Unless the pair $S_t, A_t$ appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
Append $G$ to $Returns(S_t, A_t)$
$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)
For all $a \in \mathcal{A}(S_t)$:
$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

Policy Improvement Theorem: (for any $s \in S$)

$$q_\pi(s, \underbrace{\pi'(s)}_{\text{greedy}}) = \sum_a \pi'(a|s)\, q_\pi(s,a)$$

$$= \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \max_a q_\pi(s,a)$$

$$\geq \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \sum_a \frac{\pi(a|s) - \frac{\varepsilon}{|\mathcal{A}(s)|}}{(1-\varepsilon)} q_\pi(s,a)$$

$$= \frac{\varepsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) - \frac{\varepsilon}{|\mathcal{A}(s)|} \sum q_\pi(s,a) + \sum_a \pi(a|s)\, q_\pi(s,a)$$

$$= v_\pi(s)$$

$$\therefore v_{\pi'}(s) \geq v_\pi(s) \quad \forall s \in \mathcal{S}$$

[

$$q_\pi(s,a) \leq \max q_\pi(s,a)$$

$$\sum \frac{\pi(a|s) - \frac{\varepsilon}{|\mathcal{A}(s)|}}{(1-\varepsilon)} q_\pi(s,a) \leq \max q_\pi(s,a) \sum_a \frac{\pi(a|s) - \varepsilon/|\mathcal{A}(s)|}{(1-\varepsilon)}$$

$$= \frac{1}{(1-\varepsilon)} - \frac{\varepsilon}{(1-\varepsilon)}$$

$$= \max q_\pi(s,a)$$

]

- consider a new env, same action, state set

$$(s,a) \begin{cases} \xrightarrow{1-\varepsilon} \text{like old env} \\ \xrightarrow{\varepsilon} \text{repicks action at random } a' \end{cases}$$

$(s,a')$ behave like old

$\tilde{v}_*$, $\tilde{q}_*$

$q_*$

$\pi$ optimal among $\varepsilon$-soft iff $v_\pi = \tilde{v}_*$

$$\tilde{v}_*(s) = \max_a \sum_{s',r} \left[ (1-\varepsilon)\, p\left(\frac{s',r}{s,a}\right) + \sum_{a'} \frac{\varepsilon}{|A(s)|}\, p\left(\frac{s',r}{s,a'}\right) \right] \left[ r + \gamma v'_*(s) \right]$$

$$= (1-\varepsilon) \max_a \sum p\left(\frac{s',r}{s,a}\right) \left[ r + \gamma \tilde{v}_*(s') \right]$$

$$+ \frac{\varepsilon}{|A(s)|} \sum_a \sum_{s'r} p\left(\frac{s'r}{s\,a}\right) \left[ r + \gamma \tilde{v}_*(s') \right]$$

when $\varepsilon$-soft policy $\pi$ can no longer be improved

$$v_\pi(s) = (1-\varepsilon) \max_a q_\pi(s,a) + \frac{\varepsilon}{|A(s)|} \sum q_\pi(s,a)$$

$$= (1-\varepsilon) \max_a \sum_{s',r} p\left(\frac{s'r}{s\,a}\right) \left[ r + \gamma v_\pi(s') \right]$$

$$+ \frac{\varepsilon}{|A(s)|} \sum_a \sum_{s'r} p\left(\frac{s'r}{s\,a}\right) \left[ r + \gamma v_\pi(s') \right]$$