Strojové učenie II

prednáška 4 – Monte Carlo metódy

Ing. Ján Magyar, PhD.
Katedra kybernetiky a umelej inteligencie
Technická univerzita v Košiciach

2021/2022 letný semester

Monte Carlo učenie posilňovaním

- prístup založený na odhade hodnotových funkcií
- model-free – nepotrebuje úplnú znalosť prostredia, spolieha sa na skúsenosti s ním
- interakcia s prostredím
  - skutočná
  - simulovaná
- obmedzenie na epizodické úlohy
- inkrementálny v zmysle epizóda po epizóde

Monte Carlo odhad $v_\pi(s)$

- cieľom je získať odhad $v_\pi = E_\pi[G_t|S_t = s]$
  - z epizodických sekvencií $S_1, A_1, R_2, S_2, A_2, R_3, S_3, A_3, \ldots$
  - odhad pre stav zo sekvencie začínajúcej v danom stave
- založené na spriemerňovaní kumulatívnych odmien
  - očakávaná kumulatívna odmena je nahradená priemernou kumulatívnou odmenou
  - so zvyšujúcim sa počtom vzoriek bude priemer konvergovať k očakávanej hodnote
- výber kumulatívnej hodnoty
  - prvá návšteva
  - každá návšteva

Algoritmus MC odhadu v_π

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

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

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

Loop forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

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 $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow \text{average}(Returns(S_t))$

Zdroj: Sutton-Barto: Reinforcement Learning, 2nd ed., 2018

Odhady stavov

- odhady stavov
  - navzájom nezávislé
  - možné odhadnúť nie všetky ale iba nejakú podmnožinu vybraných stavov
- principiálny update
  - $N(s) \leftarrow N(s) + 1$
  - $R(s) \leftarrow R(s) + \ G_t$
  - $V(s) \leftarrow R(s)/N(s)$
- inkrementálny update
  - $N(s) \leftarrow N(s) + 1$
  - $V(s) \leftarrow V(s) + (G_t - V(s))/N(s)$

Výber akcií

- použitím hodnotovej funkcie stavu
  1. odvodiť hodnotovú funkciu akcie $q_\pi(s,a)$ – hľadanie do hĺbky 1
     $$\begin{aligned} q_\pi(s,a) &= E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = S, A_t = a] \\ &= E_\pi[R_{t+1} | S_t = S, A_t = a] + \gamma E_\pi[G_{t+1} | S_t = S, A_t = a] \\ &= \sum_{r \in R} p(r|s,a) \cdot r + \gamma \sum_{s' \in S} p(s'|s,a) \cdot v_\pi(s') \end{aligned}$$
  2. vybrať maximalizujúcu akciu
- nemáme model dynamiky prostredia $p(s', r|s,a)$

Odhad $q_\pi(s, a)$

- pre výber akcií je nutné odhadovať $q_\pi(s, a)$
  - dá sa upraviť MC predikcia pre estimáciu q_π namiesto v_π
  - v sekvencii sa bude sledovať výskyt páru **stav-akcia** namiesto iba stavu

Porovnanie algoritmov

odhad $\boldsymbol{V(s)}$

- $V(s) \epsilon R$
- $Returns(s)$
- unless S_t appears in $S_0, S_1, \ldots$
- append G to $Returns(S_t)$
- $V(S_t) \leftarrow average(Returns(S_t))$

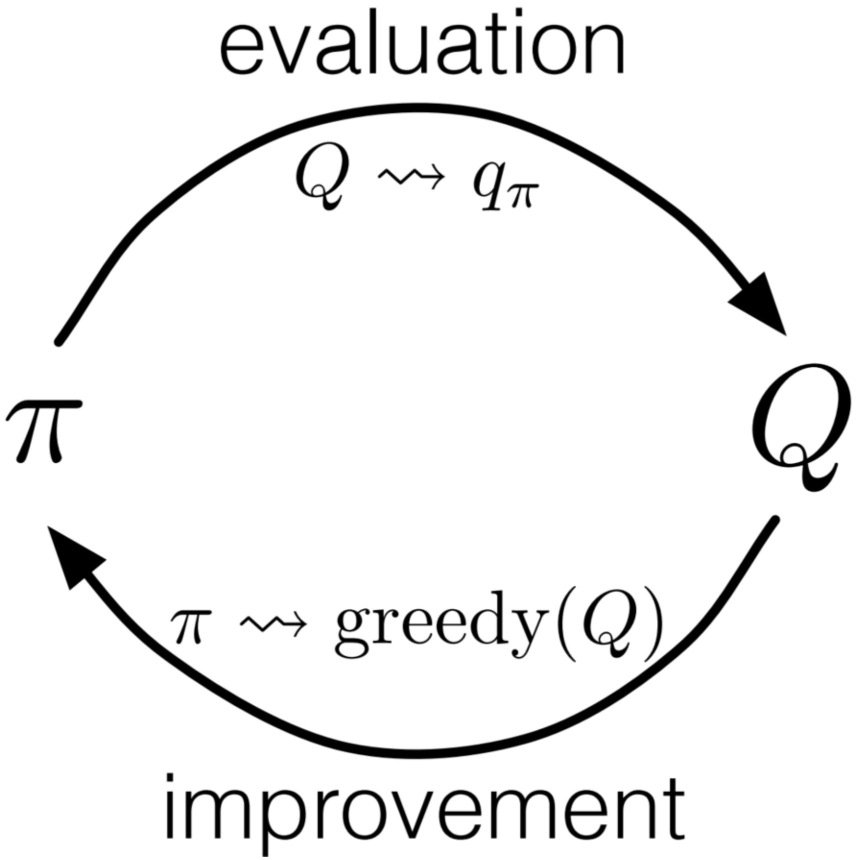
odhad $\boldsymbol{Q(s,a)}$

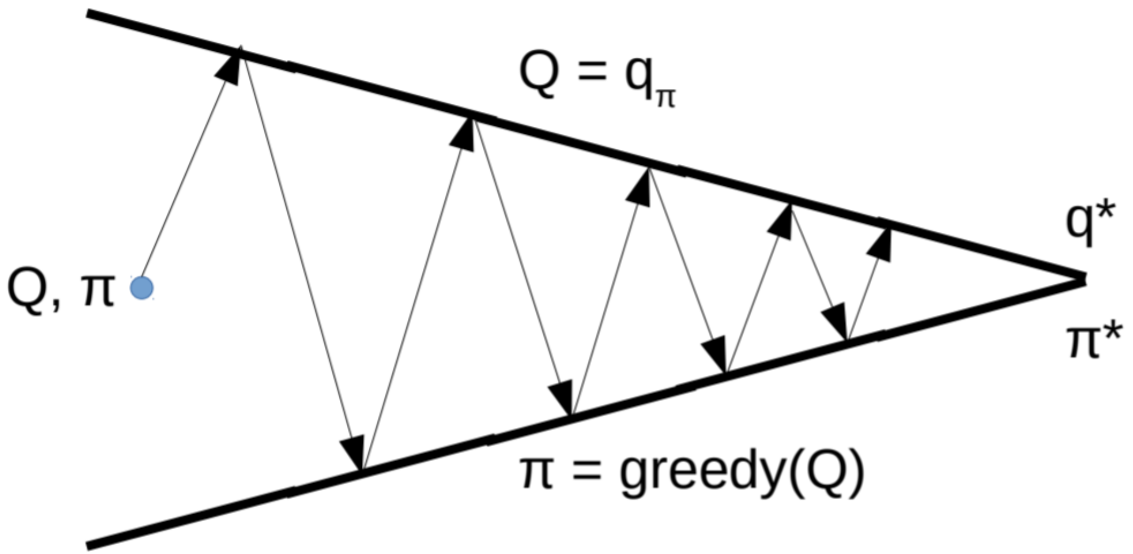
- $Q(s,a) \epsilon R$
- $Returns(s,a)$
- unless S_t, A_t appears in $S_0, A_0, S_1, A_1, \ldots$
- append G to $Returns(S_t, A_t)$
- $Q(S_t, A_t) \leftarrow average(Returns(S_t, A_t))$

Použitie politiky π

- problém s použitou politikou π
  - niektoré kombinácie stav-akcia sa nebudú v sekvenciách vyskytovať vôbec alebo iba príliš zriedkavo pre spoľahlivý odhad
  - deterministická politika v každom stave vyberie iba 1 akciu
  - problém so zachovávaním **explorácie**
- riešenie
  - exploračné štarty
    - epizóda štartuje v zadanej kombinácii stav-akcia
    - pravdepodobnosť dvojice stav-akcia začínať sekvenciu je nenulová
  - stochastické politiky
    - $p(a|s) > 0$ pre všetky stavy a k nim príslušné akcie

Všeobecná iterácia politiky





Zdroj: Sutton-Barto: Reinforcement Learning, 2nd ed., 2018

Aproximácia optimálnej politiky

$$\pi_0 \xrightarrow{E} q_{\pi_0} \xrightarrow{Z} \pi_1 \xrightarrow{E} q_{\pi_1} \xrightarrow{Z} \pi_2 \xrightarrow{E} \dots \xrightarrow{Z} \pi_* \xrightarrow{E} q_*$$

- zlepšenie politiky: $\pi(s) = \underset{a}{\operatorname{argmax}}\, q(s,a)$

$$\begin{aligned} q_{\pi_k}\big(s, \pi_{k+1}(s)\big) &= q_{\pi_k}(s, \underset{a}{\operatorname{argmax}}\, q_{\pi_k}(s,a)) \\ &= \max_a q_{\pi_k}(s,a) \\ &\geq q_{\pi_k}(s,a) \end{aligned}$$

- predpoklady garancie konvergencie
  - pokrytie všetkých dvojíc stav-akcia
  - nekonečný počet epizód
    - dostatočná aproximácia skutočných hodnôt
    - pohyb smerom k skutočným hodnotám

On-policy a off-policy vyhodnocovanie a učenie

- dva základné prístupy
  - on-policy metódy sú zamerané na politiku použitú pri tvorbe sekvencií
  - off-policy metódy sú zamerané na politiku, ktorá je rozdielna od politiky použitej pri tvorbe sekvencií
- oblasť použitia
  - vyhodnocovanie danej politiky
  - iteračné vylepšovanie danej politiky

Algoritmus MC on-policy odhadu π_*

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

Initialize:

$\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$

$Q(s, a) \in \mathbb{R}$ (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 π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

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 $Returns(S_t, A_t)$

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

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

Zdroj: Sutton-Barto: Reinforcement Learning, 2nd ed., 2018

Mäkké politiky

- mäkká politika
  - $\pi(a|s) > 0$ pre všetky $s \in S$ a $a \in A(s)$
  - môže byť posúvaná stále viac k deterministickej optimálnej politike
- ε-greedy politika
  - väčšinu času sa vyberá akcia maximalizujúca hodnotovú funkciu akcie (pravdepodobnosť *1 –* ε), občas sa vyberie náhodne vybratá akcia (pravdepodobnosť ε)
  - je príkladom ε-mäkkej politiky $\pi(a|s) \geq \frac{\varepsilon}{|A(s)|}$
  - realizovaná ako náhodný výber podľa pravdepodobnosti

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{|A(s)|} & a = \operatorname*{argmax}_a q(s,a) \\ \dfrac{\varepsilon}{|A(s)|} & pre\ ostatn\acute{e}\ a \end{cases}$$

Algoritmus MC on-policy odhadu π_*

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

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-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 π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
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 $a \in \mathcal{A}(S_t)$:

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

Zdroj: Sutton-Barto: Reinforcement Learning, 2nd ed., 2018

Konvergencia pri použití ε-greedy politiky

$$
\begin{aligned}
v_{\pi'}(s) &= \sum_a \pi'(a|s) q_\pi(s,a) \\
&= \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \max_a q_\pi(s,a) \\
&\geq \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + (1-\varepsilon) \sum_a \frac{\pi(a|s) - \frac{\varepsilon}{|A(s)|}}{1-\varepsilon} q_\pi(s,a) \\
&= \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) - \frac{\varepsilon}{|A(s)|} \sum_a q_\pi(s,a) + \sum_a \pi(a|s) q_\pi(s,a) \\
&= v_\pi(s)
\end{aligned}
$$

Off-policy prístup

- on-line dilema – snaha naučiť optimálnu politiku avšak nutnosť použiť neoptimálnu exploračnú politiku
- použitie dvoch politík pre riešenie dilemy
  - cieľová politika π (typicky deterministická)
  - exploračná politika b (môže byť ε-mäkká)
- pokrytie politík
  - $\pi(a|s) > 0 \rightarrow b(a|s) > 0$

Výhody a nevýhody

- výhody off-policy prístupu
  - všeobecnejší prístup (zahŕňa on-policy)
  - širšie použitie (učenie z pozorovania iných, znovupoužitie skúseností, viacnásobné použitie skúseností)
- nevýhody off-policy prístupu
  - pomalšia konvergencia
  - vyššia výpočtová náročnosť

Vzorkovanie podľa dôležitosti

$$S_t, A_t, S_{t+1}, A_{t+1}, S_{t+2}, \dots, S_{T-1}, A_{T-1}, S_T$$

- pravdepodobnosť výskytu sekvencie pri politike π
  $P[A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t] = \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \dots p(S_T|S_{T-1}, A_{T-1})$
  $= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)$

- pravdepodobnosť výskytu sekvencie pri politike b

$$P[A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t] = \prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)$$

Vzorkovanie podľa dôležitosti (2)

- pomer pravdepodobností

$$W_t = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)}$$

- určenie E podľa π zo sekvencie podľa b

$$v_\pi(s) = E_\pi[W_t \cdot G_t|S_t = s]$$

Off-policy odhad $v_{\pi}(s)$

$S_{t1}, \dots, S_{T1}$ $S_{t2}, \dots, S_{T2}$ $\dots$ $S_{tn}, \dots, S_{Tn}$

- $\mathcal{T}(s)$ - množina uvažovaných výskytov stavu s v množine sekvencií
- obyčajné vzorkovanie podľa dôležitosti

$$V(s) = \frac{\sum_{i \epsilon \mathcal{T}(s)} W_i \cdot G_i}{|\mathcal{T}(s)|}$$

  - bez odchýlky, variancia nie je ohraničená
- vážené vzorkovanie podľa dôležitosti

$$V(s) = \frac{\sum_{i \epsilon \mathcal{T}(s)} W_i \cdot G_i}{\sum_{i \epsilon \mathcal{T}(s)} W_i}$$

  - odchýlka konverguje asymptoticky k nule, ohraničená variancia

Inkrementálne určovanie *V(s)*

- rekurzívny vzťah

$$V_n = \frac{W_1 G_1 + \cdots + W_n G_n}{W_1 + \cdots + W_n}$$

$$= \frac{\frac{W_1 G_1 + \cdots + W_{n-1} G_{n-1}}{W_1 + \cdots + W_{n-1}} (W_1 + \cdots + W_{n-1}) + W_n G_n}{W_1 + \cdots + W_n}$$

$$= \frac{V_{n-1}(W_1 + \cdots + W_{n-1}) + W_n V_{n-1} - W_n V_{n-1} + W_n G_n}{W_1 + \cdots + W_n}$$

$$= V_{n-1} + \frac{W_n (G_n - V_{n-1})}{W_1 + \cdots + W_n} = V_{n-1} + \frac{W_n}{C_n}(G_n - V_{n-1})$$

$$C_n = C_{n-1} + W_n$$

- inicializácia: $C_0 = 0, V_0 =$ ľubovoľná hodnota

Algoritmus MC off-policy odhadu $q_\pi(s,a)$

Off-policy MC prediction (policy evaluation) for estimating $Q \approx q_\pi$

Input: an arbitrary target policy π
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \in \mathbb{R}$ (arbitrarily)
    $C(s,a) \leftarrow 0$

Loop forever (for each episode):
    $b \leftarrow$ any policy with coverage of π
    Generate an episode following b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$, while $W \neq 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} \left[G - Q(S_t, A_t)\right]$
        $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

Zdroj: Sutton-Barto: Reinforcement Learning, 2nd ed., 2018

Off-policy odhad π_*

- použité politiky
  - cieľová: deterministická $\pi(s) \leftarrow \underset{a}{\operatorname{argmax}}\, Q(s,a)$
  - exploračná: stochastická ε-mäkká
- relatívna pravdepodobnosť vykonania kroku v sekvencii $\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$ môže byť:
  - 0 (ak $\pi(A_t|S_t) = 0$) – exploračná politika zvolila iný, ako maximalizujúci krok
  - $\frac{1}{b(A_t|S_t)}$ (ak $\pi(A_t|S_t) = 1$) – exploračná politika zvolila maximalizujúci krok
- nevýhody
  - algoritmus sa učí iba z koncových častí sekvencií jednotlivých epizód
  - pomalé učenie pri dlhých epizódach

Algoritmus MC off-policy odhadu π_*

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)
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(s,a)$ (with ties broken consistently)

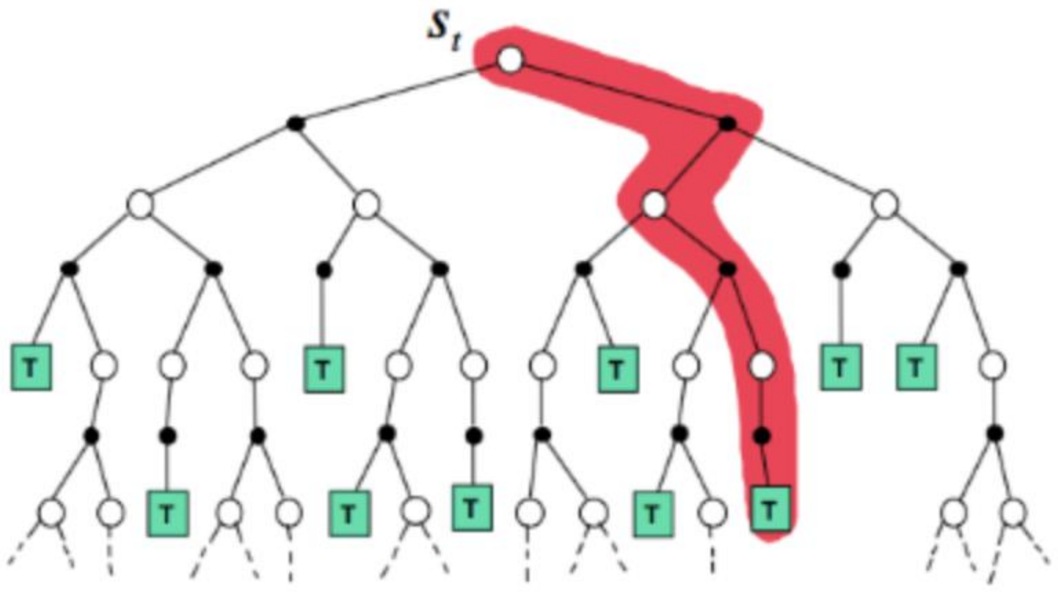
Loop forever (for each episode):
    $b \leftarrow$ any soft policy
    Generate an episode using b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t,A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
        $W \leftarrow W \frac{1}{b(A_t|S_t)}$

Zdroj: Sutton-Barto: Reinforcement Learning, 2nd ed., 2018

Backup diagram

Monte-Carlo

$$V(S_t) \leftarrow V(S_t) + \alpha\,(G_t - V(S_t))$$



Zdroj: https://lilianweng.github.io/lil-log/2018/02/19/a-long-peek-into-reinforcement-learning.html