Temporal-difference metódy
(Strojové učenie II)

M. Mach

Katedra kybernetiky a umelej inteligencie, FEI, TUKE

február 2021 - marec 2024

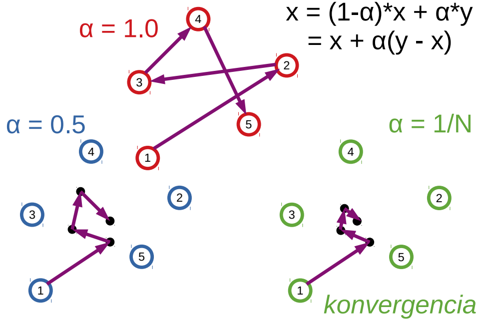
Temporal-difference učenie posilňovaním

- Cieľom je odhad hodnotových funkcií
- Kombinácia dvoch princípov
  - bootstrapping - odhad na základe iného odhadu
    (DP: $v_\pi(s) = E[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$)
  - model-free - stačí skúsenosť s prostredím (MC: netreba úplnú znalosť prostredia v zmysle modelu)
    (MC: $v_\pi(s) = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s]$)
- Interakcia s prostredím
  - skutočná
  - simulovaná
- Schopný pracovať s
  - neúplnými epizódami
  - kontinuálnymi úlohami
- Inkrementálny v zmysle krok po kroku - online

Dávkový vs. iteratívny update

$$\bar{x} \leftarrow \sum_{i=1}^{N} x_i \qquad \bar{x} \leftarrow \bar{x} + \alpha[x_i - \bar{x}] \;\; i = 1, \ldots, N$$

$$\begin{aligned} x &= (1-\alpha)*x + \alpha*y \\ &= x + \alpha(y - x) \end{aligned}$$



$$\begin{aligned}
\bar{x}_{1..5} &= \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} \\
&= \frac{\frac{x_1+x_2+x_3+x_4}{4} 4 + x_5}{5} \\
&= \frac{\bar{x}_{1..4} 4 + \color{blue}{\bar{x}_{1..4} - \bar{x}_{1..4}} + x_5}{5} \\
&= \frac{\bar{x}_{1..4} 5 + x_5 - \bar{x}_{1..4}}{5} \\
&= \bar{x}_{1..4} + \frac{1}{5}\left(x_5 - \bar{x}_{1..4}\right)
\end{aligned}$$

Rozdiel v princípoch odhadu

$$\begin{aligned} v_\pi(s) &= E_\pi[G_t|S_t = s] \\ &= E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] \end{aligned}$$

- MC
  - cieli na skutočnú kumulatívnu odmenu
  - $V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$
    $\alpha = 1/N$ (MC), $\alpha = const$ (constant-α MC)
  - odhad iba na základe skúsenosti
  - čaká na znalosť G_t (epizóda musí dobehnúť)
- TD
  - cieli na odhad kumulatívnej odmeny
  - $V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$
  - odhad na základe skúsenosti a iného odhadu (G_{t+1})
  - čaká iba jeden krok

TD chyba

- Člen v zátvorke: $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$
  - chyba - rozdiel medzi aktuálnym odhadom hodnoty pre stav S_t ($V(S_t)$) a lepším odhadom pomocou okamžitej skúsenosti (R_{t+1}) a odhadu pre nasledujúci stav ($V(S_{t+1})$)
- Ak by sa hodnoty odhadov V nemenili počas epizódy ale iba po nej

$$
\begin{aligned}
G_t - V(S_t) &= R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1}) \\
&= \delta_t + \gamma(G_{t+1} - V(S_{t+1})) \\
&= \delta_t + \gamma(R_{t+2} + \gamma G_{t+2} - V(S_{t+1}) + \gamma V(S_{t+2}) - \gamma V(S_{t+2})) \\
&= \delta_t + \gamma \delta_{t+1} + \gamma^2 (G_{t+2} - V(S_{t+2})) \\
&= \sum_{k=t}^{T-1} \gamma^{k-t} \delta_k
\end{aligned}
$$

- TD robí korekciu voči menšej chybe než MC

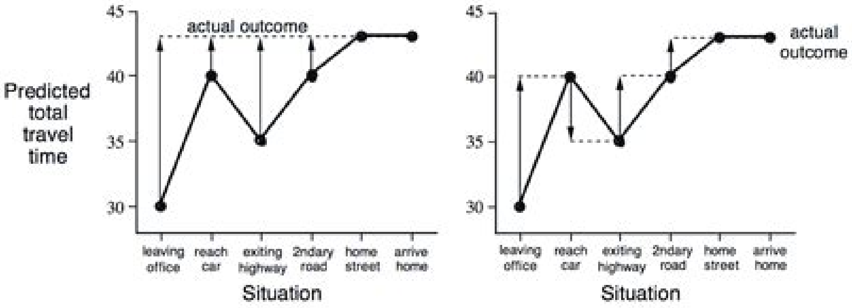
Algoritmus TD odhadu v_π

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
        $S \leftarrow S'$
    until S is terminal

Ilustračný príklad: cesta domov



- Aktualizované hodnoty - odhad po prvom dni (pre $\alpha = 1$)
  - TD - bezprostredne môže updatovať ako reakciu na aktuálnu situáciu (zmeny obomi smermi)
  - MC - musí čakať až na príchod domov (zmeny na rovnakú hodnotu)

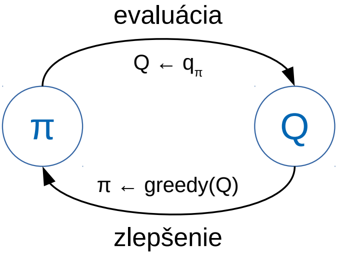
Konvergencia

$$V(S_t) \leftarrow V(S_t) + \alpha[Target - V(S_t)]$$

- Podmienky pre konvergenciu V k v_π
  - $\sum_{n=1}^{\infty} \alpha_n = \infty$
  - $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$
- TD - konverguje iba približne (v priemere)
  - α je konštantné a "dostatočne" malé (nesplnená druhá podmienka)
- MC - konverguje (s pravdepodobnosťou 1)
  - $\alpha = 1/N$ sa postupne zmenšuje (splnené sú obe podmienky)
- Asymtotická konvergencia môže byť v praxi pomalá
  - čo konverguje rýchlejšie (lepšie využije dáta)?
    - TD (offline - dávkový update) $<$ constant-α MC
    - TD (online) ? MC

Učenie politiky

$$\ldots, S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, \ldots$$



- Založené na všeobecnej iterácii politiky
- Použitie $Q(s, a)$ ako odhadu $q_\pi(s, a)$
  - neznalosť modelu
  - nestačí poznať v_π
- Problém explorácie
  - pokrytie párov (s,a)
  - ϵ-mäkká politika
- On-policy ($\pi = b$) vs off-policy ($\pi \neq b$)

Aktualizácia odhadu Q

- Spôsoby odhadu G_{t+1} v
  $q_\pi(s, a) = E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$

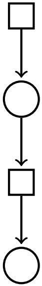
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

$$\begin{aligned} Q(S_t, A_t) \leftarrow & \\ & Q(S_t, A_t) + \alpha\,[R_{t+1} + \gamma E[Q(S_{t+1}, A_{t+1}) | S_{t+1}] - Q(S_t, A_t)] = \\ & Q(S_t, A_t) + \alpha\,[R_{t+1} + \gamma \textstyle\sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t)] \end{aligned}$$

Aktualizácia: $\ldots + \gamma Q(S_{t+1}, A_{t+1}) - \ldots$

$$\ldots, R_t, S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, \ldots$$



- $Q(S_t, A_t)$ je
  - pre neterminálny stav aktualizované podľa pätice vybranej zo sekvencie
  - nulové pre terminálny stav
- On-policy odhad
- Konvergencia (ku q_* a π_*)
  - všetky páry (s,a) navštívené nekonečnekrát
  - politika konverguje ku greedy politike ($\epsilon = 1/t$)
- Sarsa algoritmus

Algoritmus Sarsa

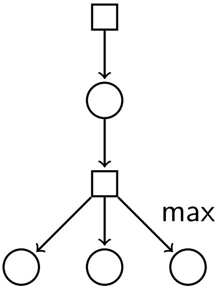
Sarsa (on-policy TD control) for estimating $Q \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Loop for each step of episode:
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma Q(S', A') - Q(S, A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

Aktualizácia: $\ldots + \gamma \max_a Q(S_{t+1}, A_{t+1}) - \ldots$

$$\ldots, R_t, S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, \ldots$$



- $Q(S_t, A_t)$ je
  - pre neterminálny stav aktualizované podľa štvorice vybranej zo sekvencie
  - nulové pre terminálny stav
- Off-policy odhad
- Konvergencia (ku q_* a π_*)
  - potrebné je zabezpečiť aktualizáciu všetkých párov (s,a)
- Q-learning algoritmus

Algoritmus Q-Learning

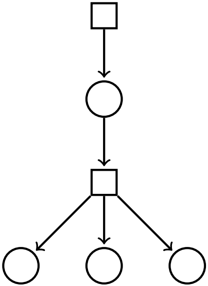
Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        Choose A from S using policy derived from Q (e.g., ε-greedy)
        Take action A, observe R, S'
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
        $S \leftarrow S'$
    until S is terminal

Aktualizácia: $\ldots + \gamma E[Q(S_{t+1}, A_{t+1})|S_{t+1}] - \ldots$

$\ldots, R_t, S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, \ldots$

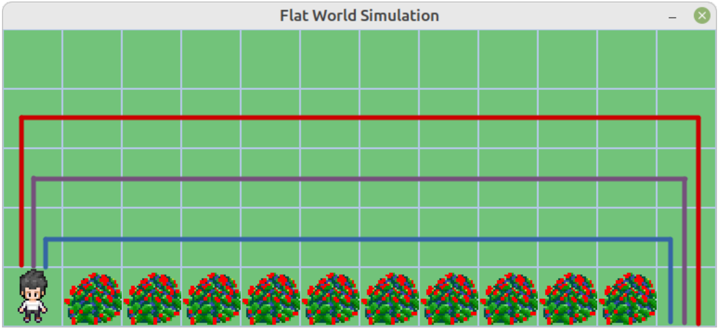


- $Q(S_t, A_t)$ je
  - pre neterminálny stav aktualizované podľa štvorice vybranej zo sekvencie
  - nulové pre terminálny stav
- Spriemerňovanie akcií v S_{t+1}
  $E[Q(S_{t+1}, A_{t+1})|S_{t+1}] = \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a)$
- On-policy/Off-policy odhad
  - π sa môže lišiť od exploračnej politiky
  - ak π je greedy, tak je Q-learning
- Expected Sarsa algoritmus

(algoritmus podobný ako pre Q-Learnig, zmena iba v jednom riadku)

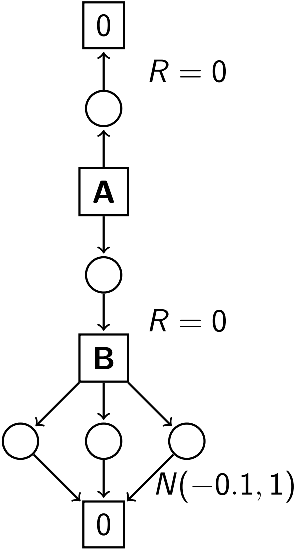
Porovnanie algoritmov

- Koniec epizódy: krík, pravé dolné políčko
- Odmena: -100 (krík), -1 (ostatné políčka)
- Parametre: 500 epizód, $\alpha = 0.2$, $\gamma = 0.9$, $\epsilon = 0.1$



Sarsa, Expected Sarsa, Q-Learning

Dvojité učenie



- Maximalizačná odchýlka
  - $E[R \in N(-0.1, 1)] < 0$
  - $max[R \in N(-0.1, 1)] > 0$
- Použitie jedného odhadu Q
  - pre určenie najlepšej akcie A_{t+1} v stave S_{t+1}
  - pre odhad hodnoty $q(S_{t+1}, A_{t+1})$
  - rozhodnutia sú závislé
- Použitie dvoch nezávislých odhadov Q_1 a Q_2
  - $Q_2(S_{t+1}, \arg\max_a Q_1(S_{t+1}, a))$
  - $Q_1(S_{t+1}, \arg\max_a Q_2(S_{t+1}, a))$
  - update iba Q_1 alebo Q_2
  - striedanie rolí Q_1 a Q_2

Algoritmus Dvojitý Q-Learning

Double Q-learning, for estimating $Q_1 \approx Q_2 \approx q_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q_1(s, a)$ and $Q_2(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, such that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using the policy ε-greedy in $Q_1 + Q_2$
    Take action A, observe R, S'
    With 0.5 probabilility:

$$Q_1(S, A) \leftarrow Q_1(S, A) + \alpha\Big(R + \gamma Q_2\big(S', \arg\max_a Q_1(S', a)\big) - Q_1(S, A)\Big)$$

    else:

$$Q_2(S, A) \leftarrow Q_2(S, A) + \alpha\Big(R + \gamma Q_1\big(S', \arg\max_a Q_2(S', a)\big) - Q_2(S, A)\Big)$$

    $S \leftarrow S'$
  until S is terminal

Dvojitá (Expected) Sarsa

$$\pi(a|s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A(s)|} & a = \arg\max_a (Q_1(s,a) + Q_2(s,a)) \\ \frac{\epsilon}{|A(s)|} & inak \end{cases}$$

Sarsa

$$Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) + \alpha[R_{t+1} + \gamma Q_2(S_{t+1}, A_{t+1}) - Q_1(S_t, A_t)]$$
$$Q_2(S_t, A_t) \leftarrow Q_2(S_t, A_t) + \alpha[R_{t+1} + \gamma Q_1(S_{t+1}, A_{t+1}) - Q_2(S_t, A_t)]$$

Expected Sarsa

$$Q_1(S_t, A_t) \leftarrow$$
$$Q_1(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q_2(S_{t+1}, a) - Q_1(S_t, A_t)\right]$$
$$Q_2(S_t, A_t) \leftarrow$$
$$Q_2(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q_1(S_{t+1}, a) - Q_2(S_t, A_t)\right]$$

Greedy exploračná politika

- Princíp - monotónny update odhadu Q
  - všetky $Q(s, a)$ inicializovať na vysoké hodnoty
  - update iba v tvare zníženia odhadu
  - explorácia
    - greedy politika v stave S_t vyberie akciu A_t
    - aktualizuje sa (zníži) hodnota $Q(S_t, A_t)$
    - hodnota pre nejakú inú akciu v S_t ostane vyššou ako pre akciu A_t
    - pri ďalšej návšteve stavu S_t greedy politika vyberie inú akciu a nie A_t
- Podmienky realizácie
  - odmena $R_i \in [0, 1]$ pre $i = 1, 2, \ldots$
  - maximálne možná kumulatívna odmena

$$G_t = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_k \leq \sum_{k=t+1}^{\infty} \gamma^{k-t-1} R_{max}$$

$$= \sum_{k=t+1}^{\infty} \gamma^{k-t-1} = \sum_{k=1}^{\infty} \gamma^{k-1} = \frac{1}{1-\gamma}$$

Oneskorený Q-Learning

- Hodnoty Q inicializované na $1/(1-\gamma)$
- Update $Q(s,a)$ sa deje na základe väčšieho množstva dát
  - nech v čase t v stave $S_t = s$ bola akcia $A_t = a$
  - ak pre (s,a) je už k dispozícii zozbieraný dostatok dát z aktuálnej a minulých návštev $(r_1,\ s'_1,\ \ldots,\ r_m,\ s'_m)$, tak
    - určí sa nová možná hodnota pre $Q_{t+1}(s,a)$

      $$U(s,a) = \frac{1}{m}\sum_{i=1}^{m}(r_i + \gamma \max_{a'} Q_t(s'_i, a')) + \epsilon_1$$

      kde úlohou ϵ_1 je napomôcť $Q(s,a) \geq Q_*(s,a)$
    - ak $Q_t(s,a) - U(s,a) \geq \epsilon_1$ tak sa update realizuje $Q_{t+1}(s,a) = U(s,a)$, inak nie
    - zozbierané dáta pre (s,a) sa zahodia
  - ak dát pre (s,a) nie je dostatok, pokračuje sa v ich zbieraní

Q-Learnig vs Oneskorený Q-Learnig

- Q-Learning: $Q_{t+1} = (1-\alpha)Q_t + \alpha(r + \gamma max_a Q(s', a))$
- Pre (s, a) bolo získané r_1 a s_1 $(\alpha = 1)$

  $Q_{t+1}(s, a) = (1-1)Q_t + (r_1 + \gamma max_{a'} Q(s_1, a')) = (r_1 + \gamma max_{a'} Q(s_1, a'))$
- Pre (s, a) bolo získané r_2 a s_2 $(\alpha = 1/2)$

  $Q_{t+2}(s, a) = (1-1/2)Q_{t+1}(s, a) + 1/2(r_2 + \gamma max_{a'} Q(s_2, a'))$

  $= 1/2(r_1 + \gamma max_{a'} Q(s_1, a')) + 1/2(r_2 + \gamma max_{a'} Q(s_2, a'))$

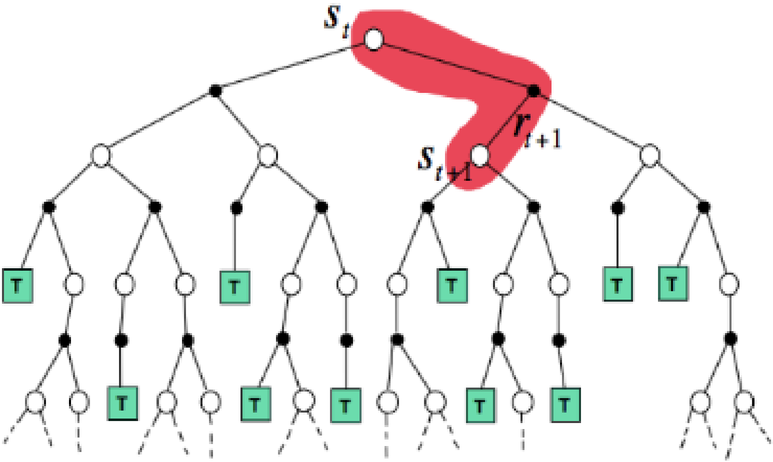
  $= 1/2 \sum_{i=1}^{2}(r_i + \gamma max_{a'} Q(s_i, a'))$
- Pre (s, a) bolo získané r_3 a s_3 $(\alpha = 1/3)$

  $Q_{t+3}(s, a) = (1-1/3)Q_{t+2}(s, a) + 1/3(r_3 + \gamma max_{a'} Q(s_3, a'))$

  $= 2/3(1/2 \sum_{i=1}^{2}(r_i + \gamma max_{a'} Q(s_i, a'))) + 1/3(r_3 + \gamma max_{a'} Q(s_3, a'))$

  $= 1/3 \sum_{i=1}^{3}(r_i + \gamma max_{a'} Q(s_i, a'))$
- Oneskorený Q-Learning = Q-Learning so zmenšujúcou sa hodnotou α (s dávkovým updatom)

Backup diagram



©/*lilianweng.github.io*