

Strojové učenie II

prednáška 3 – Dynamické programovanie

Ing. Ján Magyar, PhD.
Katedra kybernetiky a umelej inteligencie
Technická univerzita v Košiciach
2021/2022 letný semester

Dynamické programovanie

- metóda riešenia zložitých problémov
  - rekurzívne rozdelenie problému na jednoduchšie podproblémy
  - vyriešenie jednotlivých podproblémov
  - skombinovanie riešení podproblémov do riešenia problému
- vlastnosti riešeného problému
  - prekrývajúce sa podproblémy
    - riešenia podproblémov sú viacnásobne používané (cache)
    - riešenia podproblémov sú memoizované (caching)
  - optimálna štruktúra
    - optimálne riešenie celkového problému pozostáva z optimálnych riešení podproblémov

Dynamické programovanie a MDP

- rozklad na podproblémy je reprezentovaný Bellmanovými rovnicami

$$v_\pi(s) = E[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s]$$
$$q_\pi(s,a) = E[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]$$

pre rekurzívnu dekompozíciu na dve zložky

- nasledujúci krok
- ostávajúce kroky

Podúlohy v MDP

- hodnotové funkcie *v(s)* a *q(s,a)*
  - používané pre určenie hodnoty všetkých predchodcov (stavov a dvojíc stav-akcia)
  - ukladané v tabuľke pre opakované použitie
  - na základe $v_*(s)$ a $q_*(s,a)$ sa určí optimálna politika

Dynamické programovanie pre RL

- počítanie optimálnej politiky na základe perfektného modelu prostredia (niekedy označované ako plánovanie)
  - model v tvare MDP
  - konečný MDP
  - dynamika procesu daná distribúciou $p(s', r|s, a)$
- použitie pre úlohy
  - diskrétny priestor stavov a akcií – možné riešiť priamo
  - spojitý priestor – priamo kvantovaním alebo zložitejšími prístupmi
- limitované využitie
  - predpoklad dostupnosti perfektného modelu
  - veľké výpočtové nároky

Vyhodnocovanie politiky π

- označuje určovanie hodnotových funkcií v_π a q_π pre ľubovoľnú politiku π

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)(r + \gamma v_\pi(s'))$$

- sústava lineárnych rovníc
  - počet rovníc daný počtom možných stavov
- garancia jedinečného riešenia ak platí aspoň jedno:
  - $\gamma < 1$
  - z každého stavu je pri π dosiahnutý terminálny stav

Iteračné vyhodnocovanie politiky π

- iteračné riešenie sústavy rovníc
  - $v_0(s) = 0$ pre terminálne stavy
  - počiatočná aproximácia pre neterminálne stavy $v_0(s)$ môže byť ľubovoľná
  - je generovaná sekvencia aproximácií hodnotovej funkcie stavu $v_0, v_1, v_2, \ldots$
- aktualizačné pravidlo

$$v_{k+1}(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)(r + \gamma v_k(s'))$$

- ukončenie ak $\max_s |v_{k+1}(s) - v_k(s)|$ je dostatočne malé

Iteračné vyhodnocovanie politiky π

- aktualizačné stratégie
  - sweep stratégia
    - dve polia, jedno pre nové hodnoty v_{k+1} a jedno pre staré hodnoty v_k
    - nové hodnoty počítané iba zo starých hodnôt
    - staré hodnoty sa počas výpočtu nových hodnôt nemenia
  - in-place stratégia
    - jedno pole reprezentujúce staré aj nové hodnoty
    - vypočítaná nová hodnota okamžite prepíše starú hodnotu
    - nové hodnoty počítané zo starých aj nových hodnôt

Iteračné vyhodnocovanie politiky π

- konvergencia
  - $\{v_k\}$ konverguje k v_π ak $k \to \infty$
  - in-place
    - konverguje rýchlejšie ako verzia s dvomi poliami
    - rýchlosť konvergencie je ovplyvnená poradím aktualizácie hodnôt stavov

Algoritmus odhadu v_π

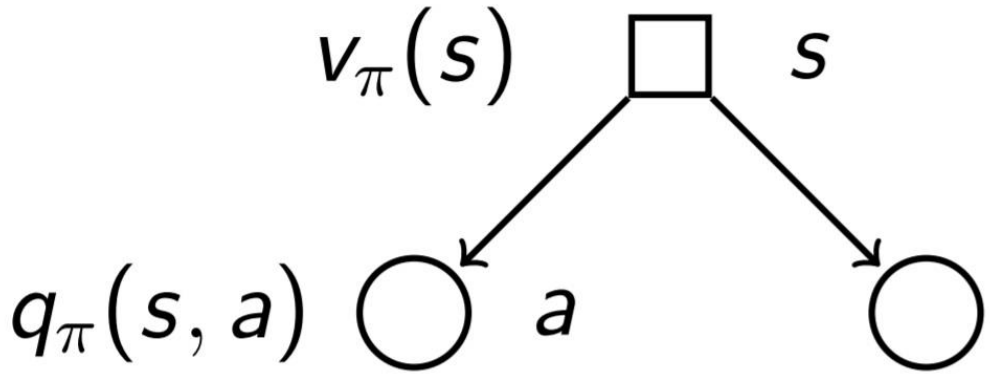
Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a) \left[r + \gamma V(s') \right]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

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

Hľadanie lepšieho výberu akcie



- vzťah medzi $v_\pi(s)$ a $q_\pi(s,a)$

$$\min_a q_\pi(s,a) \leq v_\pi(s) \leq \max_a q_\pi(s,a)$$

- nech pre stav s: $v_\pi(s) < q_\pi(s,a)$
  - $v_\pi(s)$ – ako dobre je zo stavu s pokračovať podľa π
  - lepšie je v s nevyberať podľa π ale raz vybrať a a potom pokračovať podľa π
  - lepšie je v s vždy vyberať akciu a

Teoréma zlepšovania politiky

- nech π a π' sú deterministické politiky
- nech $q_\pi(s, \pi'(s)) \geq v_\pi(s)$ pre všetky $s \in S$
- potom π' musí byť rovnako dobrá alebo lepšia ako π
- potom $v_{\pi'}(s) \geq v_\pi(s)$

Teoréma zlepšovania politiky

$$
\begin{aligned}
v_\pi(s) &\le q_\pi\big(s, \pi'(s)\big) \\
&= E_{\pi'}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s] \\
&\le E_{\pi'}\big[R_{t+1} + \gamma q_\pi\big(S_{t+1}, \pi'(S_{t+1})\big)\big| S_t = s\big] \\
&= E_{\pi'}[R_{t+1} + \gamma[R_{t+2} + \gamma v_\pi(S_{t+2})] | S_t = s] \\
&\le E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi\big(S_{t+2}, \pi'(S_{t+2})\big)\right]\big| S_t = s\big] \\
&= \cdots \\
&\le E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s\right] \\
&= v_{\pi'}(s)
\end{aligned}
$$

Zlepšovanie politiky

- rozšírenie TZP na všetky stavy a všetky akcie

$$\pi'(s) = \operatorname{argmax}_a q_\pi(s,a)$$

$$\pi'(s) = \operatorname{argmax}_a E[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = a]$$

$$\pi'(s) = \operatorname{argmax}_a \sum_{s'} \sum_{r} p(s', r|s, a)(r + \gamma v_\pi(s'))$$

- výber najlepšej akcie s uvážením následkov jedného dopredného kroku
- zlepšovanie politiky ako greedy výber akcií s ohľadom na hodnotovú funkciu pôvodnej politiky

Zlepšovanie politiky

- ak $a = \pi(s)$ je deterministická politika, potom môžeme zlepšiť politiku daným rozšírením na $\pi'(s)$

$$q_\pi\big(s, \pi'(s)\big) = \max_a q_\pi(s, a) \geq q_\pi\big(s, \pi(s)\big) = v_\pi(s)$$

- ak π' nepriniesla zlepšenie, tak:

$$q_\pi\big(s, \pi'(s)\big) = \max_a q_\pi(s, a) = q_\pi\big(s, \pi(s)\big) = v_\pi(s)$$

čo je Bellmanova rovnica optimality a $v_\pi(s) = v_*(s)$ a π je optimálnou politikou

Greedy politika

- greedy politika spĺňa podmienky teorémy
  - je lepšia alebo rovnaká ako politika, podľa ktorej boli vytvorené hodnotové funkcie, na základe ktorých vznikla greedy politika
  - ak je rovnako dobrá, ako pôvodná politika, tak obe sú optimálne
- zlepšovanie politiky – proces pretvárania politiky na greedy politiku na základe hodnotových funkcií pôvodnej politiky
- rozšírenie na stochastickú politiku
  - všetky maximalizujúce akcie majú nenulovú pravdepodobnosť výberu
  - ostatné akcie majú nulovú pravdepodobnosť výberu

Iterácia politiky

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

- sekvencia monotónne sa zlepšujúcich politík a hodnotových funkcií
  - na základe politiky π a k nej prislúchajúcej hodnotovej funkcie v_π možno vytvoriť lepšiu politiku π',
  - ktorá umožní lepšiu hodnotovú funkciu $v_{\pi'}$,
  - ktorá umožní vytvoriť lepšiu politiku π'', …
- pre konečný MDP sekvencia
  - má konečný počet iterácií
  - konverguje k optimálnej politike

Algoritmus iterácie politiky

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
      $\Delta \leftarrow 0$
      Loop for each $s \in \mathcal{S}$:
         $v \leftarrow V(s)$
         $V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s))\big[r + \gamma V(s')\big]$
         $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
      old-action $\leftarrow \pi(s)$
      $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
      If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

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

Iterácia hodnôt

- evaluácia politiky nemusí skonvergovať (v rámci limitu Δ)
  - stačí menej iterácií vyhodnotenia politiky (zmeny správnym smerom aj keď ešte mimo Δ)
  - extrémnym prípadom je iba jedna iterácia
- kombinácia zlepšovania politiky a skráteného (jednokrokového) ohodnocovania
  - jeden krok aktualizácie hodnotenia politiky (funkcia *v*) zahŕňajúcej spôsob vylepšovania politiky
  - počas iterovania nie je potrebné explicitné vytváranie politiky
- sekvencia $\{v_k\}$ konverguje k v_*
  - je to pravidlo vytvorené z Bellmanovej funkcie optimality

Kombinácia zlepšovania a ohodnocovania

- jednokrokové ohodnocovanie politiky ($a = \pi(s)$)

$$v_{k+1} = \sum_{s'} \sum_{r} p(s', r|s, a)(r + \gamma v_k(s'))$$

- zlepšovanie politiky

$$\pi(s) = \operatorname{argmax}_{a} \sum_{s'} \sum_{r} p(s', r|s, a)(r + \gamma v_k(s'))$$

- kombinácia zlepšovania politiky a skráteného ohodnocovania

$$v_{k+1}(s) = \max_{a} \sum_{s'} \sum_{r} p(s', r|s, a)(r + \gamma v_k(s'))$$

Algoritmus iterácie hodnôt

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$

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

Algoritmy synchrónneho DP

Algoritmus	Zdrojová teória
iteratívne vyhodnocovanie politiky	Bellmanova rovnica očakávania
iterácia politiky	Bellmanova rovnica očakávania + greedy zlepšovanie politiky
iterácia hodnôt	Bellmanova rovnica optimality

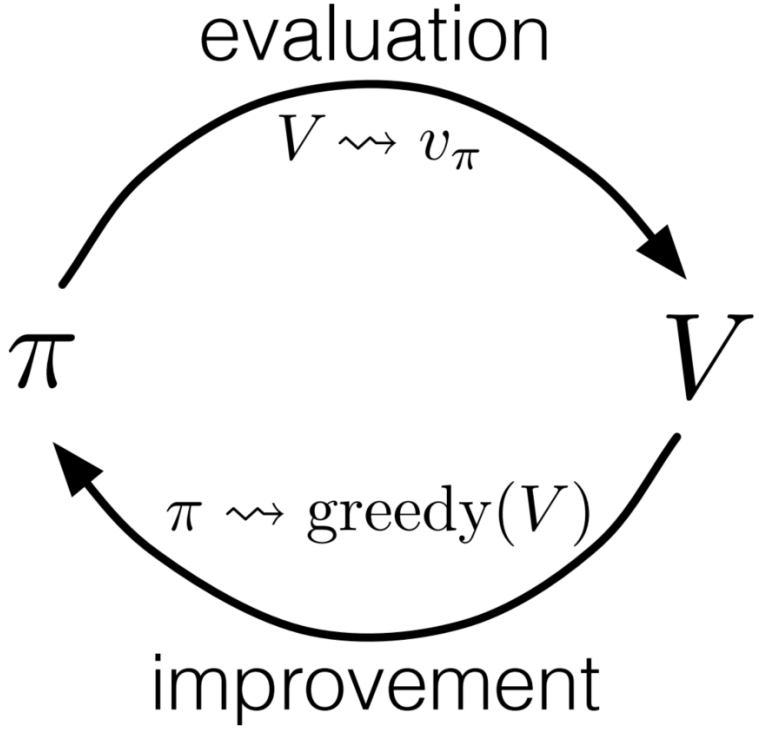
- založené na práci s hodnotovou funkciou stavu
- zložitosť je polynomiálna vzhľadom na počet stavov a počet akcií

Asynchrónne DP

- sekvencia úplných prechodov všetkých stavov je nevhodná ak
  - množina stavov je veľmi veľká
  - nie všetky stavy sú zaujímavé z pohľadu optimálnej politiky
- asynchrónne DP algoritmy
  - in-place iterácie
  - nie sú systematické prechody celou množinou stavov
    - niektoré stavy môžu byť aktualizované viackrát kým iné budú aktualizované iba raz
    - kvôli konvergencii nemožno stavy vynechať úplne
  - flexibilita ohľadom poradia aktualizácie stavov

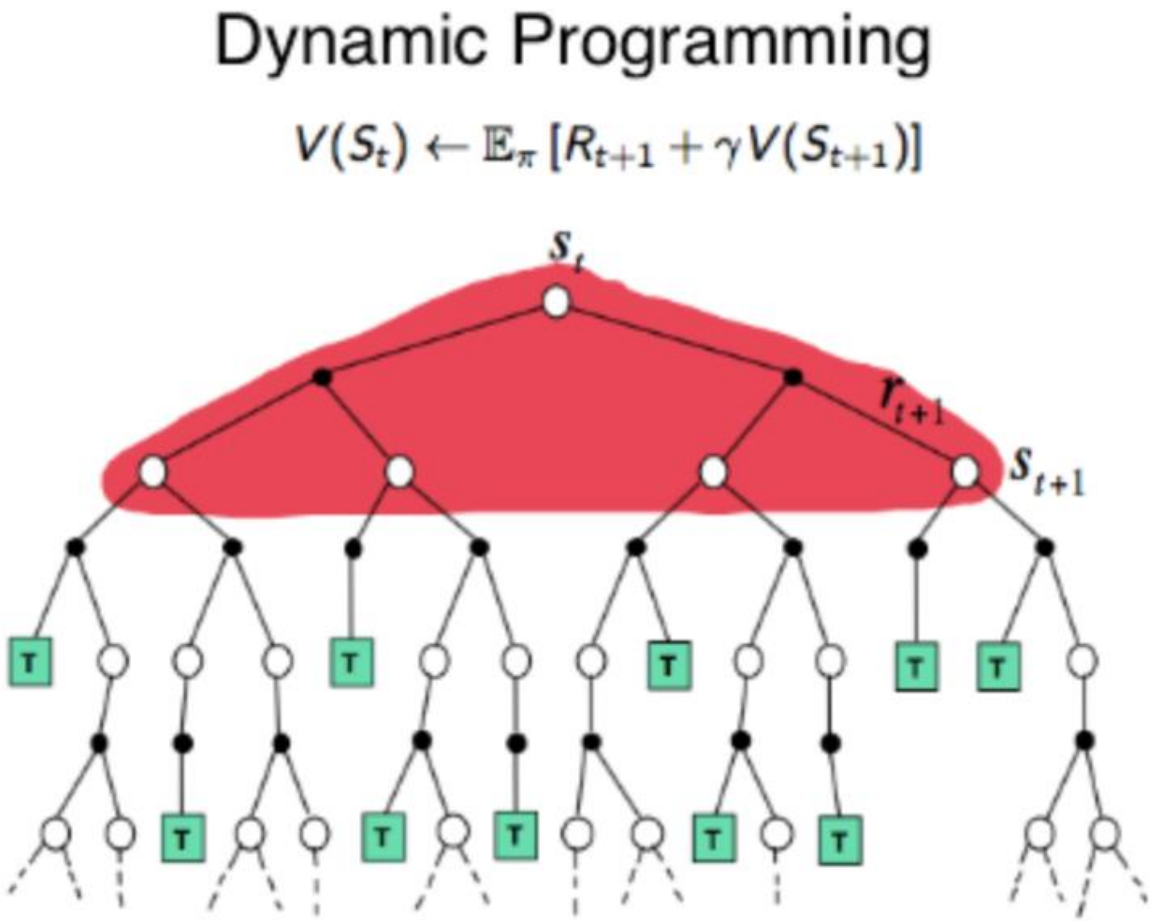
Interakcia evaluácie a zlepšovania

- interakcia dvoch procesov
  - rôzne granularity striedania
  - procesy súťažia a kooperujú
- konvergencia smerom k optimálnym v_* a π_*
  - iba spoločná stabilizácia



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

Backup diagram



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