Strojové učenie II

prednáška 7 – Aproximácia politiky

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

Hodnotové funkcie vs politika

- prístup založený na hodnotových funkciách
  - hodnotové funkcie učené: $v(s), \hat{v}(s, \overline{w}), q(s, a), \hat{q}(s, a, \overline{w})$
  - politika implicitná (ε-greedy, greedy)
- prístup založený na politike
  - hodnotové funkcie neučené
  - politika učená: $\pi(a \middle| s, \bar{\theta}) = P[A_t = a | S_t = s, \bar{\theta}_t = \bar{\theta}]$
  - *policy gradient algoritmy*
- prístup založený na hodnotových funkciách a politike
  - hodnotové funkcie učené: $v(s), \hat{v}(s, \overline{w})$
  - politika učená: $\pi(a \middle| s, \bar{\theta}) = P[A_t = a | S_t = s, \bar{\theta}_t = \bar{\theta}]$
  - *actor-critic algoritmy*

Aproximácia politiky

- parametrizovaná podoba stochastickej (diferencovateľnej) politiky $\pi(a|s,\bar{\theta}) \in (0,1)$
- preferencie akcií $h(s,a,\bar{\theta}) \in \mathcal{R}$
  - možná injektáž apriórnych znalostí
- soft-max transformácia na pravdepodobnosti
  - ľubovoľne blízke približovanie deterministickej politike

$$\pi(a|s,\bar{\theta}) = \frac{e^{h(s,a,\bar{\theta})}}{\sum_b e^{h(s,b,\bar{\theta})}}$$

- parametrizácia preferencií (aproximátor preferencií)
  - lineárna kombinácia príznakov

$$h(s,a,\bar{\theta}) = \bar{\theta}^T \bar{x}(s,a)$$

Gradientové učenie parametrov politiky $\bar{\theta}$

- nech $J(\bar{\theta})$ je skalárna miera výkonu politiky s ohľadom na parametre politiky
- cieľom je $J(\bar{\theta})$
  - hľadá sa také $\bar{\theta}$, ktoré maximalizuje $J(\bar{\theta})$
- použitie gradientu
  - aktualizačné pravidlo

    $$\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha \nabla \mathrm{J}(\bar{\theta}_t)$$

    pohyb v smere gradientu

Učenie v epizodickom prostredí

- meranie výkonu politiky $J(\bar{\theta}) = v_{\pi\theta}(s_0)$ keď s_0 je prvý (štartovací) stav epizódy
  - výkon závisí na parametroch $\bar{\theta}$ pri
    - výbere akcií (závislosť známa)
    - distribúcii stavov, v ktorých sa akcie vyberajú (závislosť zvyčajne neznáma)
- teoréma policy gradientu

$$\nabla J(\bar{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta})$$

  - kde $\propto$ je „proporciálny k“
  - nepotrebuje vyjadrovať deriváciu distribúcie stavov

Odvodenie náhrady $\nabla J(\bar{\theta})$

$$\nabla J(\bar{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta})$$

$$= E_\pi \left[\sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta}) \right]$$

$$= E_\pi \left[\frac{\pi(a|S_t,\bar{\theta})}{\pi(a|S_t,\bar{\theta})} \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta}) \right]$$

$$= E_\pi \left[\sum_a \pi(a|S_t,\bar{\theta}) q_\pi(s,a) \frac{\nabla \pi(a|s,\bar{\theta})}{\pi(a|S_t,\bar{\theta})} \right]$$

$$= E_\pi \left[E_\pi\big(q_\pi(S_t,a)\big) \frac{\nabla \pi(a|s,\bar{\theta})}{\pi(a|S_t,\bar{\theta})} \right]$$

$$= E_\pi \left[E_\pi[G_t|S_t,A_t] \frac{\nabla \pi(A_t|S_t,\bar{\theta})}{\pi(A_t|S_t,\bar{\theta})} \right] = E_\pi\left[G_t \frac{\nabla \pi(A_t|S_t,\bar{\theta})}{\pi(A_t|S_t,\bar{\theta})}\right]$$

Aktualizácia parametrov $\bar{\theta}$

- aktualizačné pravidlo

$$\begin{aligned}\bar{\theta}_{t+1} &= \bar{\theta}_t + \alpha \nabla J(\bar{\theta}_t) \\ &= \bar{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t | S_t, \bar{\theta}_t)}{\pi(A_t | S_t, \bar{\theta}_t)} \\ &= \bar{\theta}_t + \alpha G_t \nabla \ln \pi(A_t | S_t, \bar{\theta}_t)\end{aligned}$$

- intuitívna interpretácia aktualizačného pravidla
- ak soft-max v spojení s lineárnou kombináciou príznakov, tak

$$\nabla \ln \pi(A_t | S_t, \bar{\theta}_t) = \bar{x}(s, a) - \sum_b \pi(b | s, \bar{\theta}) \bar{x}(s, b)$$

Algoritmus REINFORCE

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

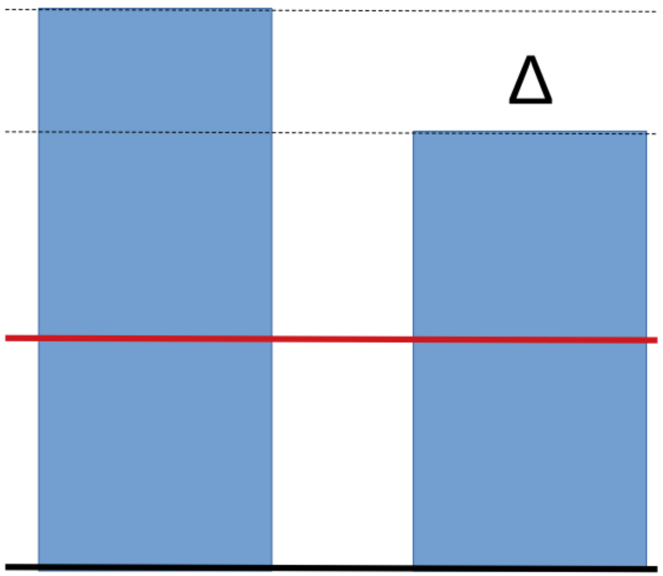
Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})$

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

Posun základne

- rozdiel
  - absolútny/relatívny
- posun nezávislý na akcii



$$\nabla J(\bar{\theta}) \propto \sum_s \mu(s) \sum_a (q_\pi(s,a) - b(s)) \nabla \pi(a|s,\bar{\theta})$$

$$\sum_a b(s) \nabla \pi(a|s,\bar{\theta}) = b(s) \nabla \sum_a \pi(a|s,\bar{\theta}) = b(s) \nabla 1 = 0$$

Algoritmus REINFORCE so základňou

REINFORCE with Baseline (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
    Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
    Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
        $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
        $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S_t,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$

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

Pridanie kritika

- REINFORCE je v podstate MC metóda
- MC → TD pre zrýchlenie konvergencie

$$\begin{aligned}
\bar{\theta}_{t+1} &= \bar{\theta}_t + \alpha \nabla J(\bar{\theta}_t) \\
&= \bar{\theta}_t + \alpha\big(G_t - \hat{v}(S_t, \bar{w})\big) \frac{\nabla \pi(A_t \middle| S_t, \bar{\theta}_t)}{\pi(A_t \middle| S_t, \bar{\theta}_t)} \\
&= \bar{\theta}_t + \alpha\big(R_{t+1} + \gamma \hat{v}(S_{t+1}, \bar{w}) - \hat{v}(S_t, \bar{w})\big) \frac{\nabla \pi(A_t \middle| S_t, \bar{\theta}_t)}{\pi(A_t \middle| S_t, \bar{\theta}_t)} \\
&= \bar{\theta}_t + \alpha \delta_t \frac{\nabla \pi(A_t \middle| S_t, \bar{\theta})}{\pi(A_t \middle| S_t, \bar{\theta}_t)}
\end{aligned}$$

Algoritmus AC

One-step Actor–Critic (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^d$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
    Initialize S (first state of episode)
    $I \leftarrow 1$
    Loop while S is not terminal (for each time step):
        $A \sim \pi(\cdot|S, \boldsymbol{\theta})$
        Take action A, observe S', R
        $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S, \boldsymbol{\theta})$
        $I \leftarrow \gamma I$
        $S \leftarrow S'$

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

Učenie v kontinuálnom prostredí

- meranie výkonu politiky $J(\bar{\theta})$ priemernou odmenou

$$J(\bar{\theta}) = r(\pi) = \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} E[R_t | S_0, A_{0:t-1} \sim \pi] =$$

$$= \sum_{s} \mu_\pi(s) \sum_{a} \pi(a|s) \sum_{r} r \sum_{s'} p(s', r | s, a)$$

kde $\mu_\pi(s)$ je ustálená distribúcia (ergodický MDP)

$$\mu_\pi(s) = \lim_{t\to\infty} P[S_t = s | A_{0:t-1} \sim \pi]$$

$$\mu_\pi(s') = \sum_{s} \mu_\pi(s) \sum_{a} \pi(a | s, \bar{\theta}) p(s' | s, a)$$

Kritik pre kontinuálne úlohy

- diferenčná odmena
$$G_t = R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \ldots$$
- diferenčná hodnotová funkcia
$$v_\pi(s) = E_\pi[G_t | S_t = s]$$
- diferenčná forma TD chyby
$$\delta_t = \left(R_{t+1} - \overline{R_t} + \hat{v}(S_{t+1}, \overline{w}_t)\right) - \hat{v}(S_t, \overline{w}_t)$$
kde $\overline{R_t}$ je odhad $r(\pi)$ v čase t
- teoréma policy gradientu ostáva naďalej platná aj pre kontinuálny prípad

Algoritmus AC (kontinuálny)

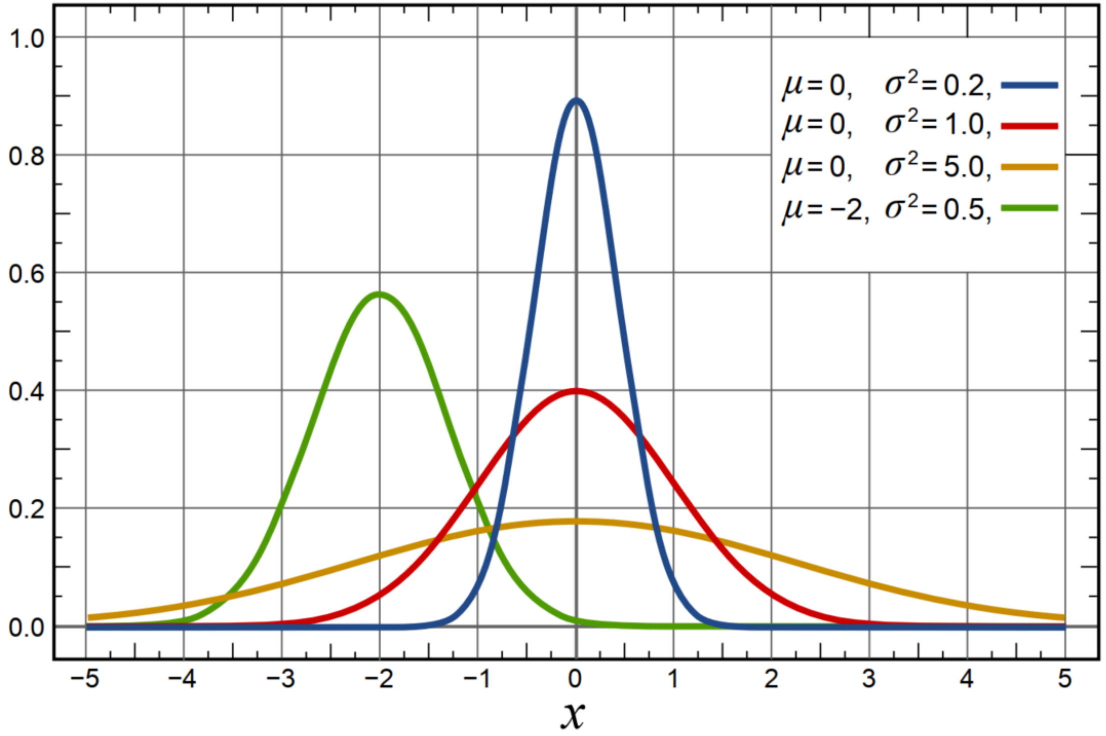
- *zmeny voči epizodickému algoritmu AC:*

Parameters: step sizes $\alpha^{\theta} > 0, \alpha^{\mathbf{w}} > 0, \alpha^{\bar{R}} > 0$
Initialize $\bar{R} \in \mathcal{R}$ (napr. na 0)
Loop forever (for each episode):

Initialize S (first state of episode)
$I \leftarrow 1$
Loop ~~while S is not terminal~~ forever (for each time step)
$\delta \leftarrow R - \bar{R} + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$
$\bar{R} \leftarrow \bar{R} + \alpha^{\bar{R}} \delta$
$\theta \leftarrow \theta + \alpha^{\theta} I \delta \nabla \ln \pi(A, S, \theta)$
$I \leftarrow \gamma I$

Spojité akcie

- akcie sú reálne skaláry
- parametrizácia politiky
  $\pi(a|s,\bar{\theta}) =$
  $$\frac{1}{\sigma(s,\bar{\theta})\sqrt{2\pi}}\exp\left(-\frac{(a-\mu(s,\bar{\theta}))^2}{2\sigma(s,\bar{\theta})^2}\right)$$
- $\bar{\theta} = \left[\bar{\theta}_\mu, \bar{\theta}_\sigma\right]^T$
- lineárne aproximátory
  $\mu(s,\bar{\theta}) = \bar{\theta}_\mu{}^T \bar{x}_\mu(s)$
  $\sigma(s,\bar{\theta}) = \exp(\bar{\theta}_\sigma{}^T \bar{x}_\sigma(s))$



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

Výhody a nevýhody metód aproximujúcich politiku

- výhody
    - lepšie konvergenčné vlastnosti
    - efektívne pre mnohorozmerný alebo spojitý priestor akcií
    - vedia učiť stochastické politiky s vhodnou úrovňou explorácie, blížiace sa deterministickým politikám
    - pre niektoré problémy je jednoduchšie parametricky reprezentovať politiku než hodnotové funkcie
- nevýhody
    - typicky konvergujú k lokálnemu a nie globálnemu optimu
    - vyhodnotenie politiky je typicky neefektívne