Algoritmy typu Actor-Critic

(Strojové učenie II)

M. Mach

Ústav umelej inteligencie, FEI, TUKE

marec 2026

Gradient politiky

- Diferencovateľný parametrický aproximátor
- Aktualizačné pravidlo pre pohyb v smere gradientu

$$\begin{aligned} \theta_{t+1} &= \theta_t + \alpha \nabla_\theta J(\pi_\theta) \\ \nabla_\theta J(\pi_\theta) &= E_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(a_t|s_t)\ \Phi_t \right] \end{aligned}$$

kde Φ_t môže byť (reinforce, reinforce so základňou, jednokrokový aktor-kritik, ...)

$$\begin{aligned} \Phi_t &= G_t = \sum_{i=t+1}^{T} \gamma^{i-t-1} R_i \\ &= G_t - v(s_t) \\ &= R_{t+1} + \gamma v(s_{t+1}) - v(s_t) \\ &= Q_\pi(s_t, a_t) \end{aligned}$$

Zisková (advantage) funkcia

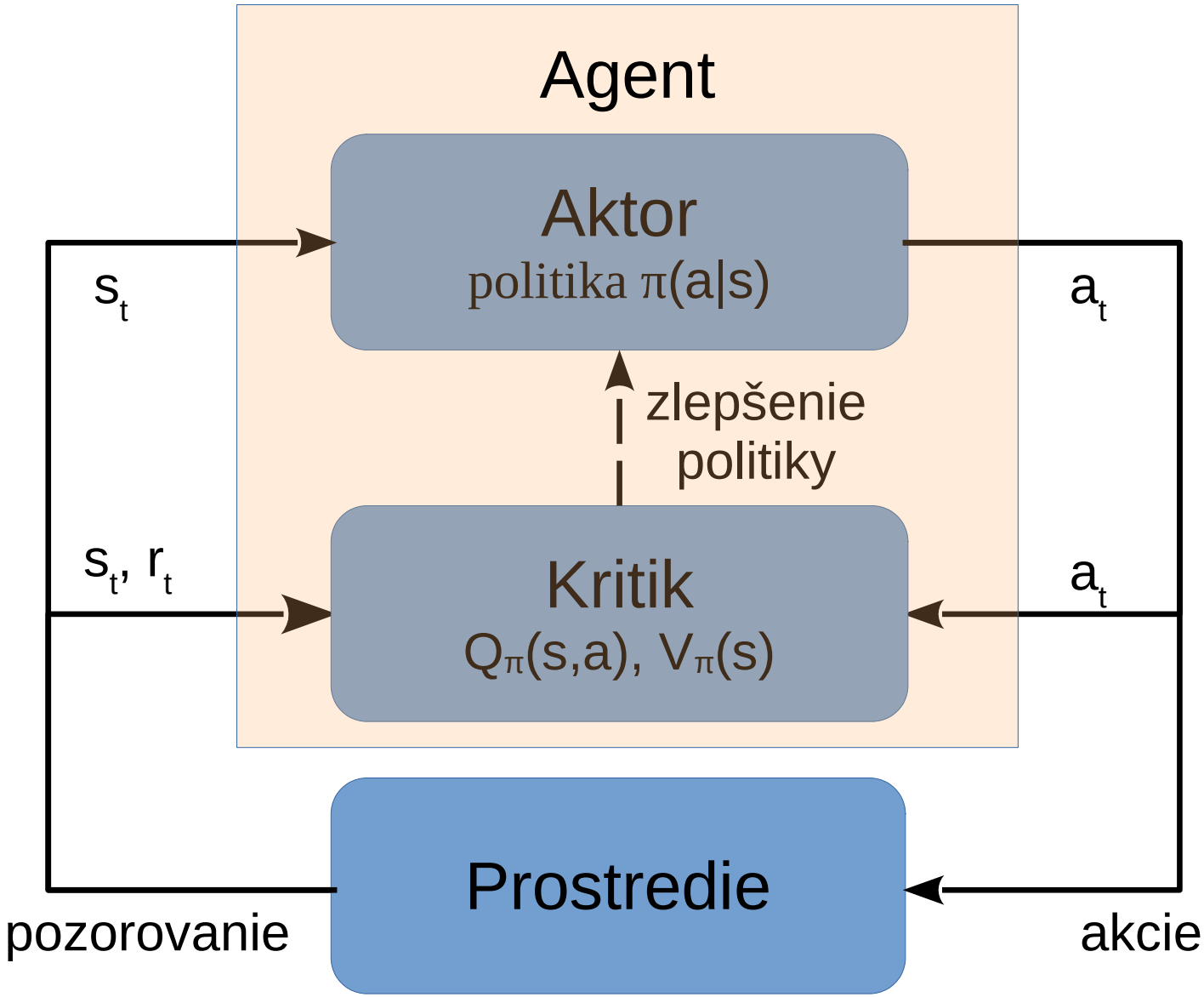
- Definícia ziskovej funkcie

$$A_\pi(s_t, a_t) \quad = \quad Q_\pi(s_t, a_t) - V_\pi(s_t)$$

- Popisuje o koľko je akcia a_t lepšia alebo horšia ako aktuálna politika daná funkciou V
  - o koľko je špecifická akcia lepšia než náhodne zvolená akcia na základe pravdepodobností $\pi(\cdot \mid s_t)$
- Ohodnocuje akciu v relatívnom zmysle
  - kladná ale aj záporná hodnota
  - ak je politika optimálna, tak iba nekladné hodnoty
- Gradient politiky

$$\begin{aligned} \nabla_\theta J(\pi_\theta) \quad &= \quad E_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\ (Q_\pi(s_t, a_t) - V_\pi(s_t))\right] \\ &= \quad E_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\ A_\pi(s_t, a_t)\right] \end{aligned}$$

Interakcia AC agenta s prostredím



Taxonómia AC algoritmov

- On-policy
  - One-step AC
  - A2C / A3C (Adavantage Actor-Critic / Asynchronous A2C)
  - TRPO (Trust Region Policy Optimization)
  - PPO (Proximal Policy Optimization)
- Off-policy
  - DDPG (Deep Deterministic Policy Gradient)
  - SAC (Soft Actor-Critic)

PPO - Surrogate objective

- Náhrada cieľovej (optimalizovanej) funkcie
  - pôvodná funkcia je výpočtovo nepraktická pre priamu optimalizáciu
  - optimalizácia náhradnej funkcie súčasne zlepšuje výkon aj podľa pôvodnej funkcie
  - náhradná funkcia je "prívetivá" pre optimalizáciu
- Pôvodná funkcia (často vedie na prílišný update)
  - gradient: $\nabla_\theta J(\pi_\theta) = E_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\ A_\pi(s_t, a_t)\right]$
  - stratová funkcia: $L^{PG}(\theta) = E_{\pi_\theta}\left[\log \pi_\theta(a_t|s_t)\ A_\pi(s_t, a_t)\right]$
- Náhradná funkcia (napr. v TRPO, PPO, ...)

$$L^{CPI}(\theta) = E_{\pi_\theta}\left[r_t(\theta) A_\pi(s_t, a_t)\right] = E_{\pi_\theta}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\ A_\pi(s_t, a_t)\right]$$

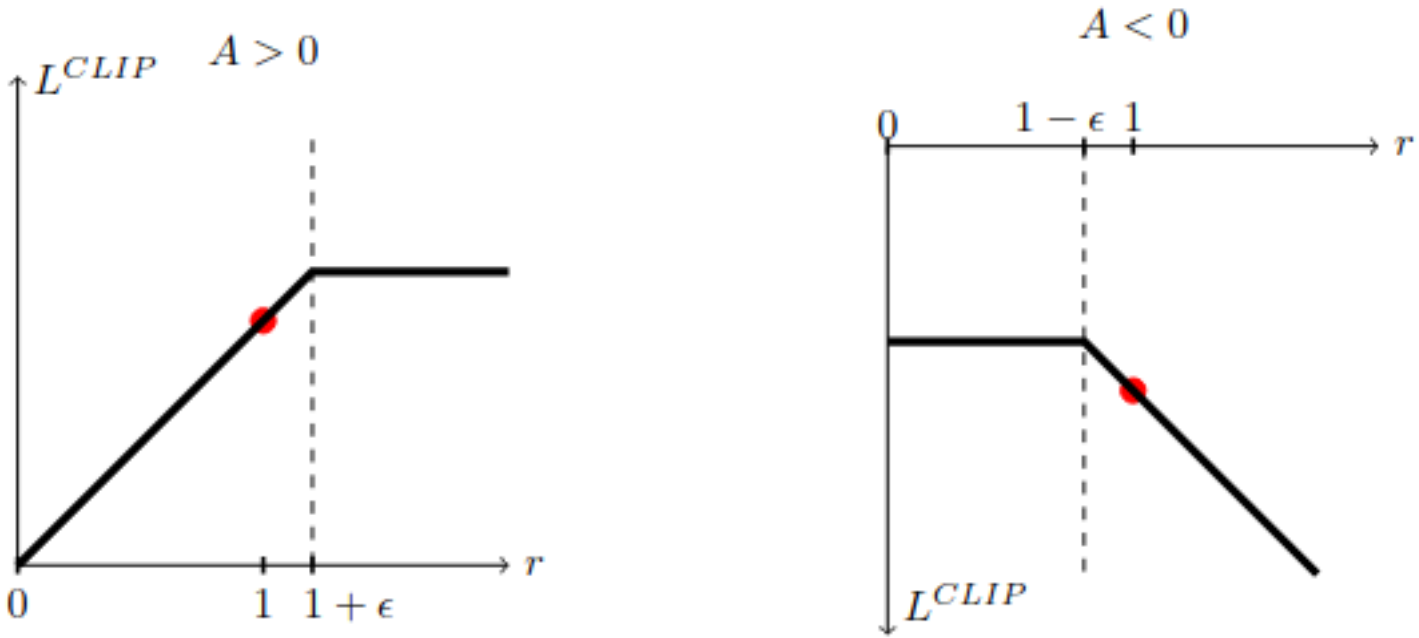
kde $r_t(\theta)$ je pravdepod. pomer $\left(r_t(\theta_{old}) = 1\right)$

Dve podoby PPO

- Typickou vlastnosťou PG metód je nestabilita
  - zmeny politiky nie sú ohraničené
  - zmeny môžu byť deštruktívne veľké
- Snaha držať updatovanú politiku blízko starej
  - nová politika v "dôveryhodnej" oblasti (zaviedlo TRPO)
- Dva základné varianty PPO
  - PPO-penalty
    - penalizácia úmerná veľkosti KL divergencie medzi starou a novou politikou
    - adaptácia penalizačného koeficientu počas učenia
  - PPO-clip
    - explicitné obmedzovanie hodnoty
    - nepoužíva KL divergenciu
    - výkonnejší ako penalizácia (podľa pôvodných autorov)
    - populárnejšia verzia
- PPO = rovnováha medzi výkonnosťou, stabilitou a jednoduchosťou

PPO - Orezanie cieľovej funkcie

$$L^{CLIP}(\theta) = E_{\pi_\theta}\left[\min(r_t(\theta)A_\pi(s_t, a_t), clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_\pi(s_t, a_t)]\right)$$



- Pesimistický odhad zlepšenia politiky
  - pozitívne zmeny obmedzuje
  - negatívne zmeny realizuje plne
- L^{CLIP} je spodné ohraničenie L^{CPI}

PPO - Zisková funkcia

- Odhad ziskovej funkcie nemôže ísť ďalej ako nejaký horizont T
- Štandardný estimátor

$$A_\pi(s_t, a_t) = -V(s_t) + r_{t+1} + \gamma r_{t+2} + \ldots + \gamma^{T-t-1} r_T + \gamma^{T-t} V(s_T)$$

- Generalizovaný estimátor

$$A_\pi(s_t, a_t) = \delta_t + (\gamma\lambda)\delta_{t+1} \ldots (\gamma\lambda)^{T-t-1}\delta_{T-1}$$

kde

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

pričom $\lambda = 1$ redukuje generalizovaný estimátor na štandardný

PPO - Štruktúra algoritmu

for iteration=1,2,... **do**
  for actor=1,2,..., N **do**
    vykonať T krokov podľa $\pi_{\theta_{old}}$
    určiť hodnoty A_1, ..., A_T
  end for
  optimalizovať L^{CLIP} s ohľadom na θ
    (K epoch, minibatch veľkosti $M \leq NT$)
  $\theta_{old} \leftarrow \theta$
end for