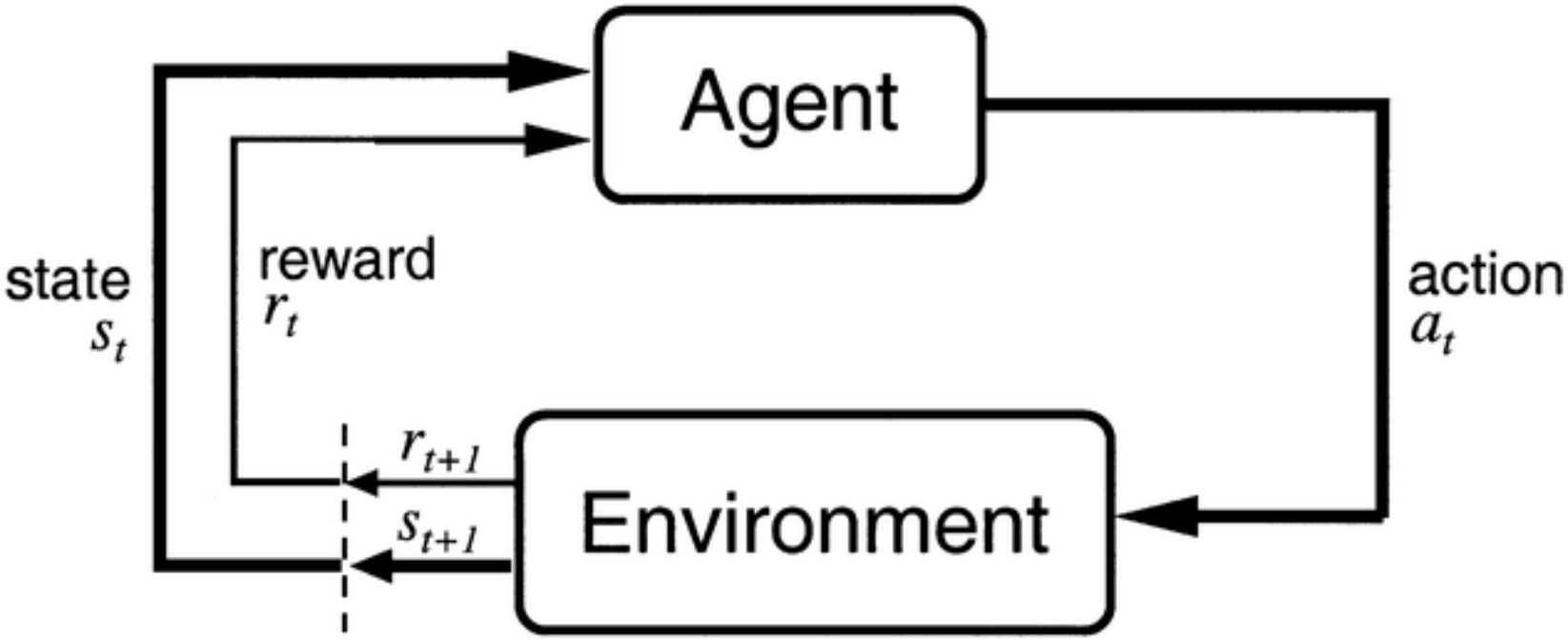
Učenie posilňovaním: hierarchia v rozhodovaní alebo vo funkčnej aproximácii?

Viliam Dillinger

Fakulta matematiky, fyziky a informatiky UK

Obsah

Agent

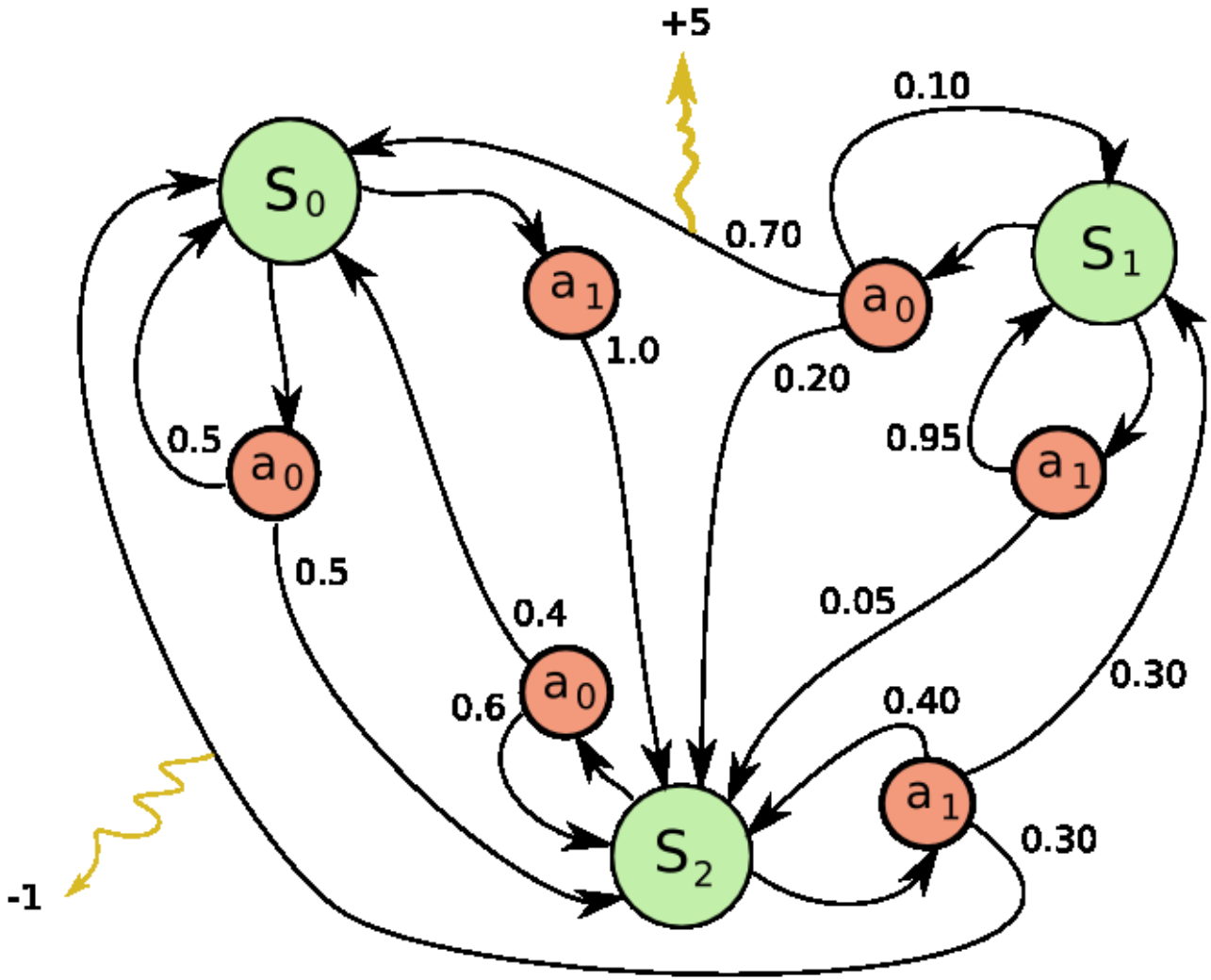


Markovov rozhodovací proces

- matematický rámec pre modelovanie rozhodovania
- štvorica $< S, A, P, R >$
  - S - konečná množina stavov
  - A - konečná množina akcií, v stave $s \in S$ môžeme vykonať A_s
  - $P(s'|s, a)$ - pravdepodobnostná prechodová funkcia
  - $R(s'|s, a)$ - funkcia odmeny
- cieľ – nájsť stratégiu $\pi(s) = a$, ktorá maximalizuje sumu diskontných odmien

  $R_t = \sum_{n=0}^{\infty} \gamma^n r_{t+n}$

Markovov rozhodovací proces



Riešenie MDP

- rozhodovanie podľa priemernej budúcej odmeny akcií v danom stave

$$\begin{aligned} V(s_t) &= r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots] = \\ &= r_t + \gamma(r_{t+1} + \gamma r_{t+2} + \ldots) = \\ &= r_t + \gamma V(s_{t+1}) \end{aligned}$$

- presný výpočet – Bellmanove rovnice, value iteration (Bellmann, 1957)

$$V(s) = \max_a \left[\sum_{s'} P(s'|s,a)(R(s'|s,a) + \gamma V(s')) \right]$$

Učenie posilňovaním

- aproximácia MDP
- trénovanie len v okolí zvolených trajektórií – lepšia aproximácia v relevantných stavoch
- nie je potrebný model prostredia (P, R)
- on-line metóda
- možnosť využitia funkčných aproximátorov
- temporal difference error (Witten 1977, Sutton a Barto 1981)

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

Najpoužívanejšie algoritmy RL

- Q-learning (off-policy) (Watkins 1989)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a'))$$

- SARSA (on-policy) (Sutton 1996)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r + \gamma Q(s',\pi(s')))$$

- Actor-critic architecture (Witten 1977, Barto et al. 1983)

$$\delta = r + \gamma V(s') - V(s)$$

$$V_{t+1}(s) = V_t(s) + \alpha_C \delta$$

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_A \delta$$

Explorácia prostredia

- nevieme zistiť, či aktuálna stratégia je optimálna
- ϵ-greedy explorácia
- Boltzmannova explorácia

$$P(\pi(s) = a) = \frac{e^{Q(s,a)/\lambda}}{\sum_{a' \in A_s} e^{Q(s,a')/\lambda}}$$

Semi-Markovov rozhodovací proces

- rozšírenie MDP – rôzne akcie môžu "trvať" rôzne dlho
- Q-učenie pre SMDP

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r + \gamma^{\tau} \max_{a'} Q(s',a'))$$

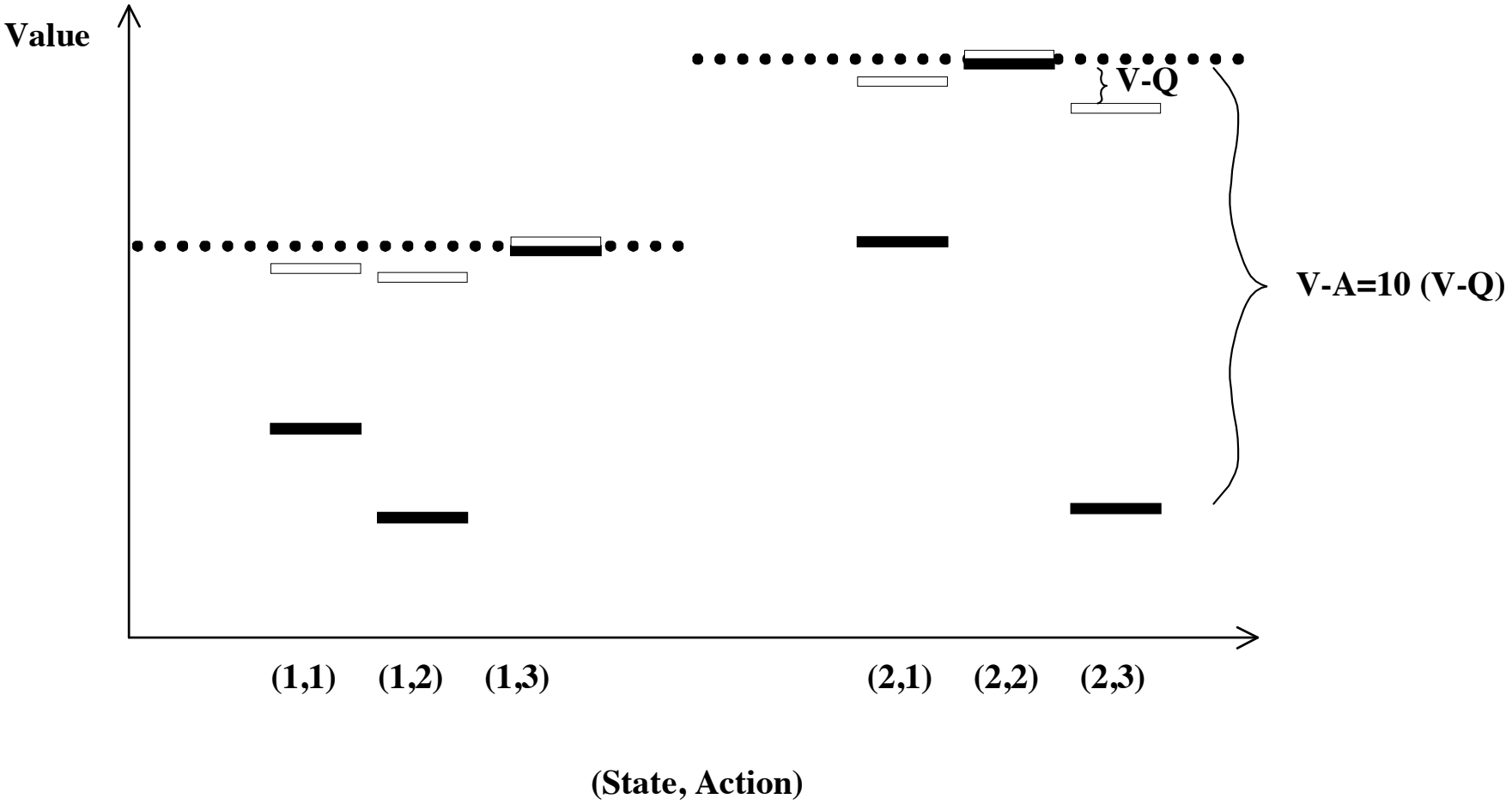
$$r = r_t + \gamma r_{t+1} + \ldots + \gamma^{t+\tau-1} r_{t+\tau}$$

Advantage learning (Baird III 1999)

- rozšírenie Q-učenia
- malá chyba pri odhade Q môže spôsobiť veľké zmeny v stratégii

$$Q(s,a) = (1-\alpha_t)Q(s,a) + \alpha_t \left(\max_{a \in A_s} Q(s,a) + \frac{r + \gamma^\tau \max_{a' \in A_{s'}} Q(s',a') - \max_{a \in A_s} Q(s,a)}{k\tau} \right)$$

Advantage learning (Baird III 1999)



- pre $1/k\tau = 10$

Viacvrstvový perceptrón (MLP)

- univerzálny funkčný aproximátor
- dopredný prechod

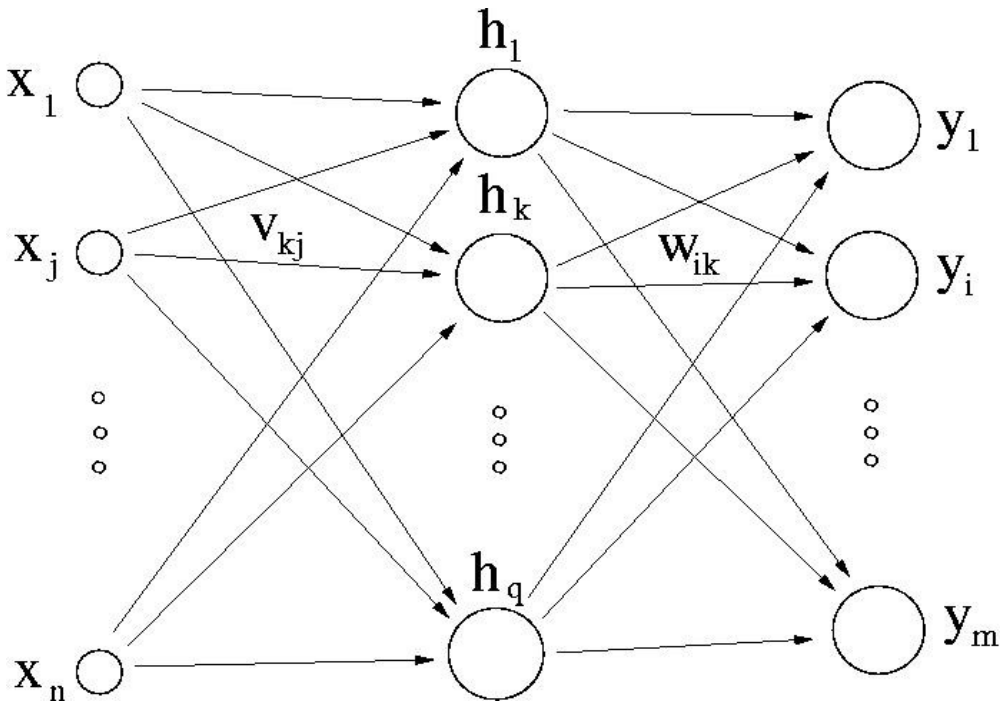
$$\mathbf{h} = f(\mathbf{V}\mathbf{x}) \quad f(net) = \frac{1}{1+e^{-net}}$$
$$\mathbf{y} = \mathbf{W}\mathbf{h}$$

- spätné šírenie chyby

$$\begin{aligned} \boldsymbol{\delta}_y &= \mathbf{d} - \mathbf{y} \\ \boldsymbol{\delta}_h &= \mathbf{W}^T \boldsymbol{\delta}_y \; .* \; \mathbf{h} \\ &\quad .* \; (1 - \mathbf{h}) \end{aligned}$$

- úprava váh

$$\mathbf{W} = \mathbf{W} + \alpha \boldsymbol{\delta}_y \mathbf{h}^T$$
$$\mathbf{V} = \mathbf{V} + \alpha \boldsymbol{\delta}_h \mathbf{x}^T$$



RBF siete

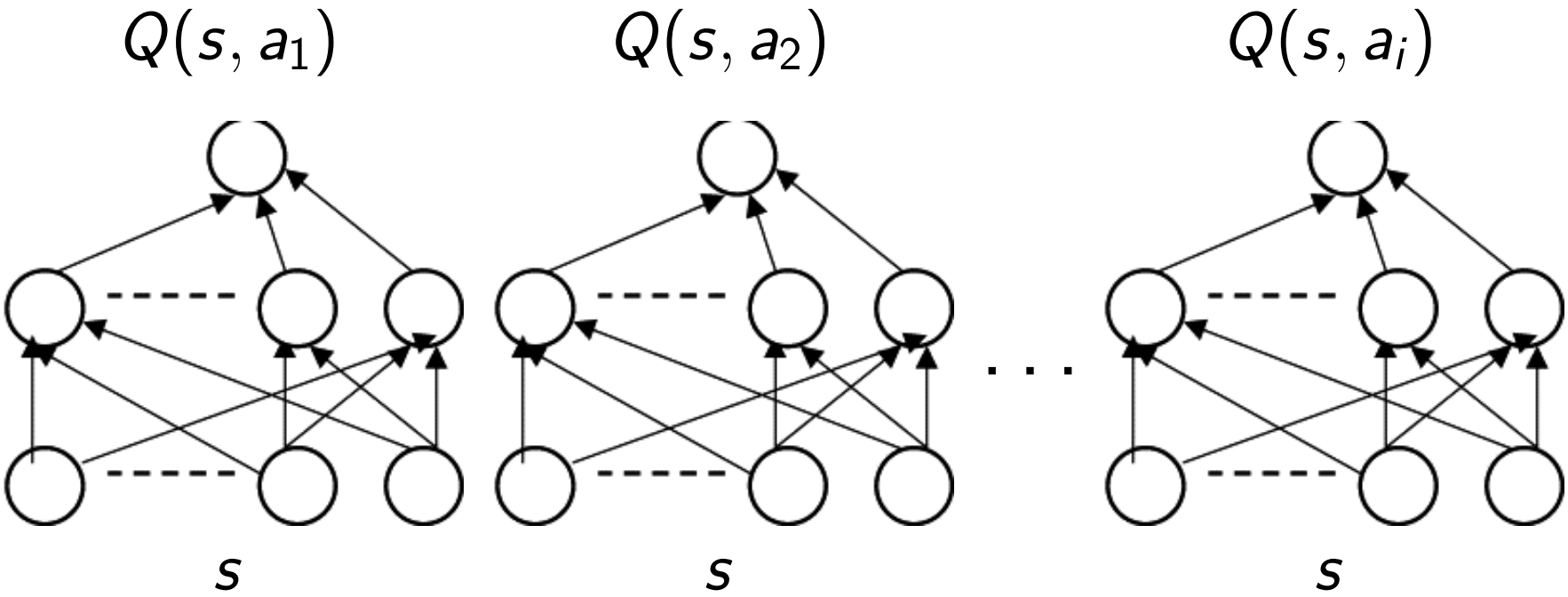
- aktivácia RBF neurónu

$$h_i = \exp(-\sigma \|\mathbf{x} - \mathbf{c}_i\|^2)$$

- pri učení sú zmeny lokálne
- potrebné väčšie množstvo neurónov (oproti MLP)

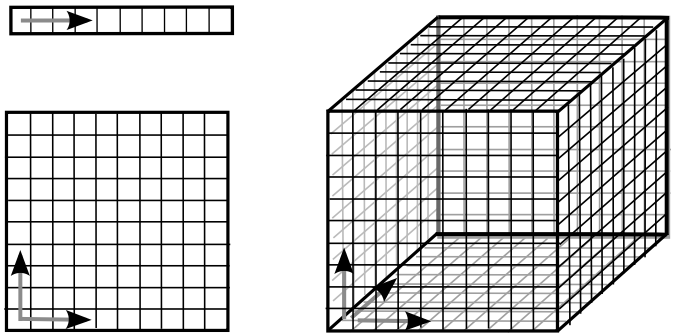
Neurónové siete a RL

- reprezentácia Q-funkcie, stratégie
- znižuje pamäťové nároky, generalizácia
- spojité prostredie

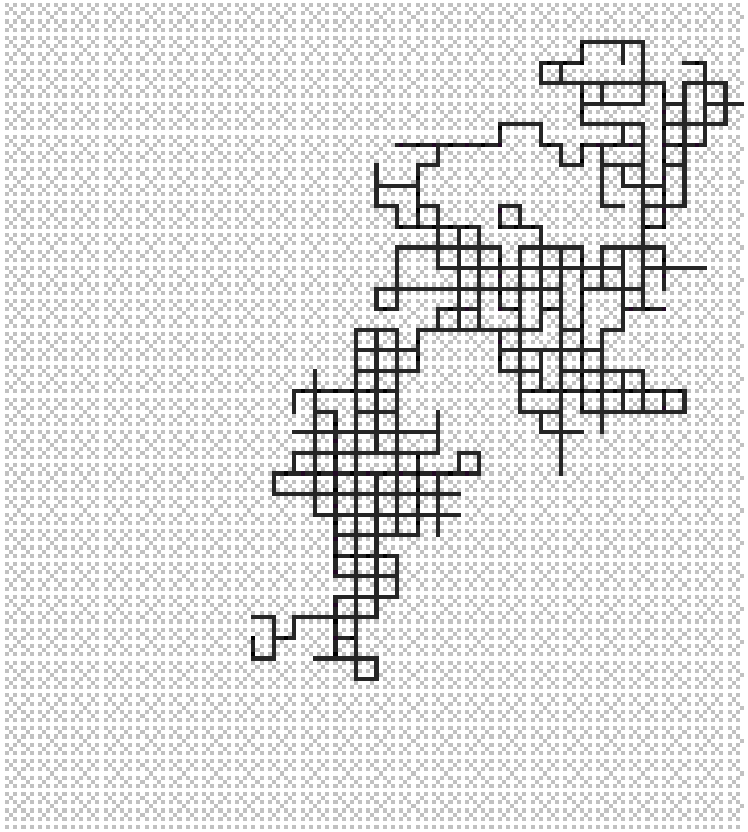


Problémy „klasického" učenia posilňovaním

- prekliatie dimenzionality
- náhodná chôdza
- šírenie odmeny
- prenos vedomostí

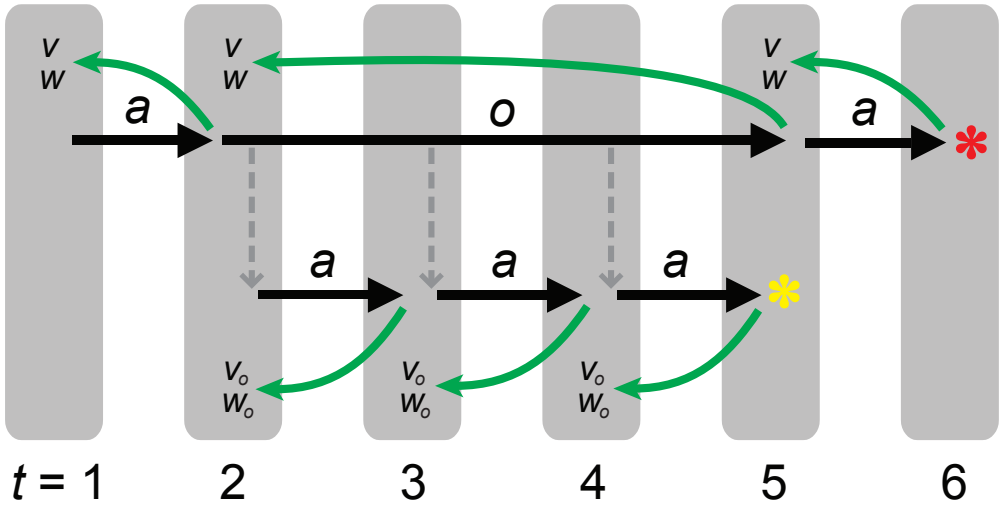


[Moerman 2009]



Hierarchické učenie posilňovaním

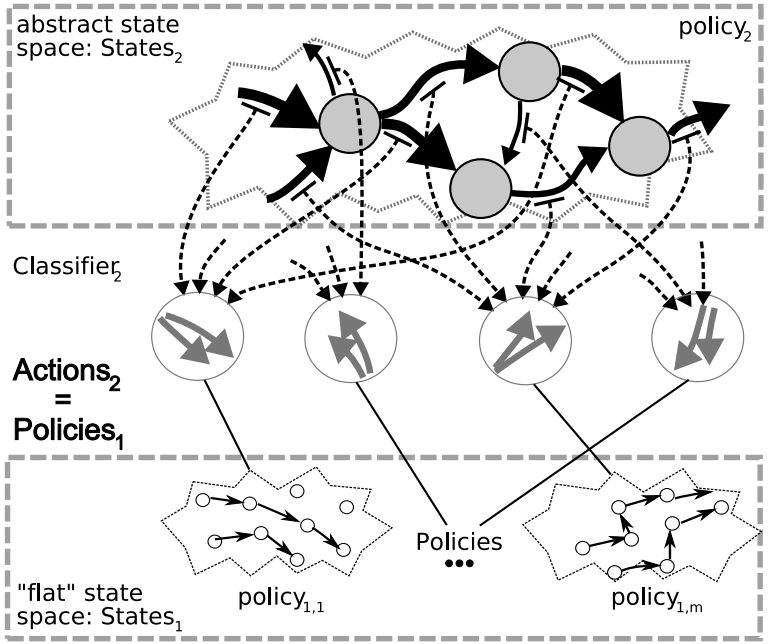
- časovo rozšírené akcie
  - prístupy využívajúce opcie (options)
  - prístupy využívajúce vrstvy
- hierarchia nad stavmi aj akciami
- hierarchicky optimálna vs optimálna

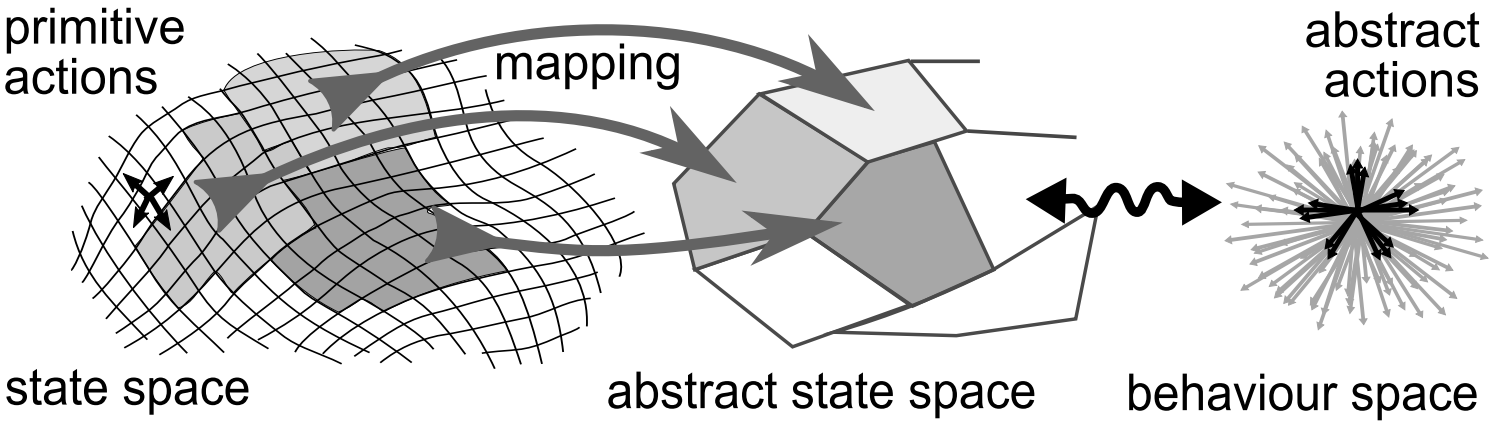


[Botvinick et. al. 2009]

HABS

- Hierarchical Assignment of Behaviours by Self-organizing [Moerman 2009]





HABS

repeat
- $reward_{HL} = 0$;
- $Policy_{HL}$ selects $SubPolicy\ SUB_i$;
- **repeat**
  - SUB_i selects and executes a primitive action ;
  - $reward_{HL} \leftarrow reward_{HL} + receivedReward$;
  - **if** *new abstract state* **then** BREAK ;
  - **else** update SUB_i with 0 ;
- **until** $timeout_{SUB}$
- **if** $timeout_{SUB}$ **then** punish SUB_i ;
- **else**
  - **if** $EXEC \in CLUSTER_{SUB}$ **then**
    - reward SUB_i ;
    - move $CLUSTER_{SUB}$ towards $EXEC$;
  - **else** punish SUB_i ;
- update $Policy_{HL}$ with $reward_{HL}$;

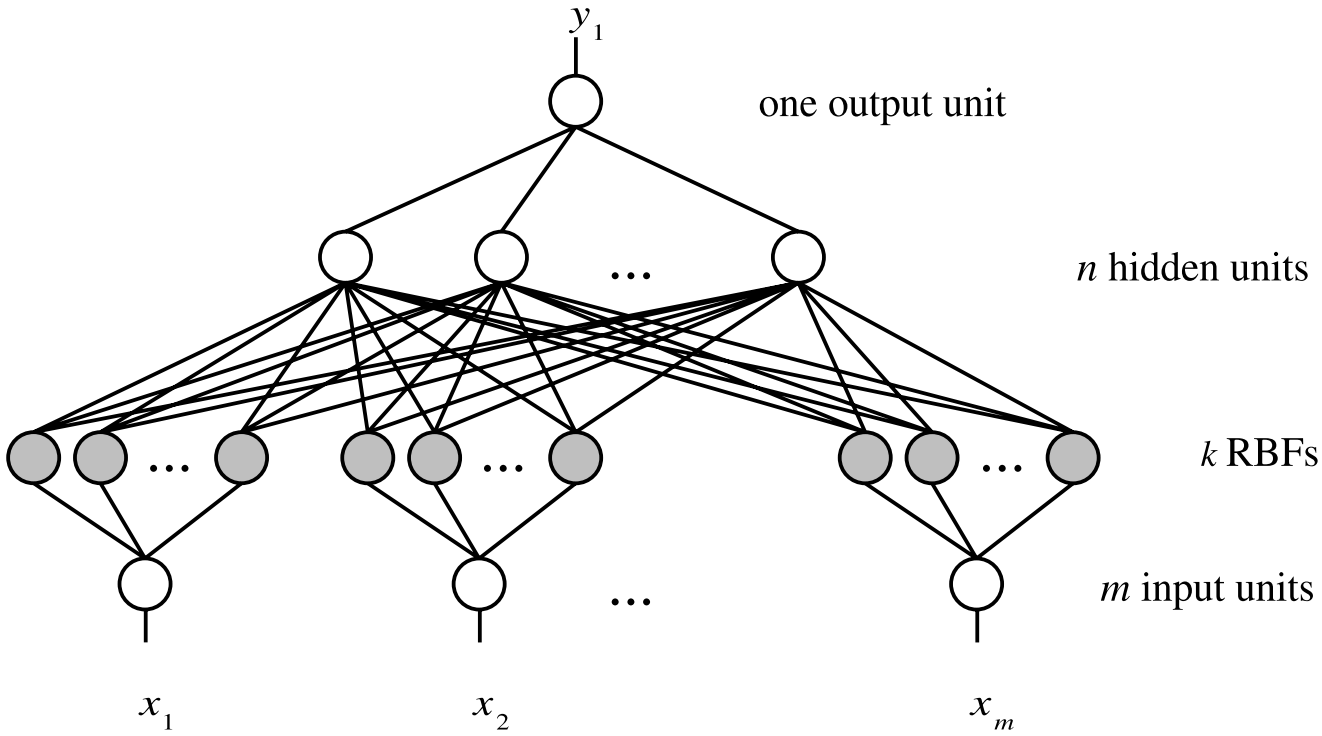
until *task solved or* $timeout_{HL}$

Naše úpravy HABS

- zrušený trest za timeout
- zmena samoorganizácie správaní – smer miesto vektora
- úprava učiaceho pravidla stratégie na vyššej vrstve – autor nevyužíval SMDP

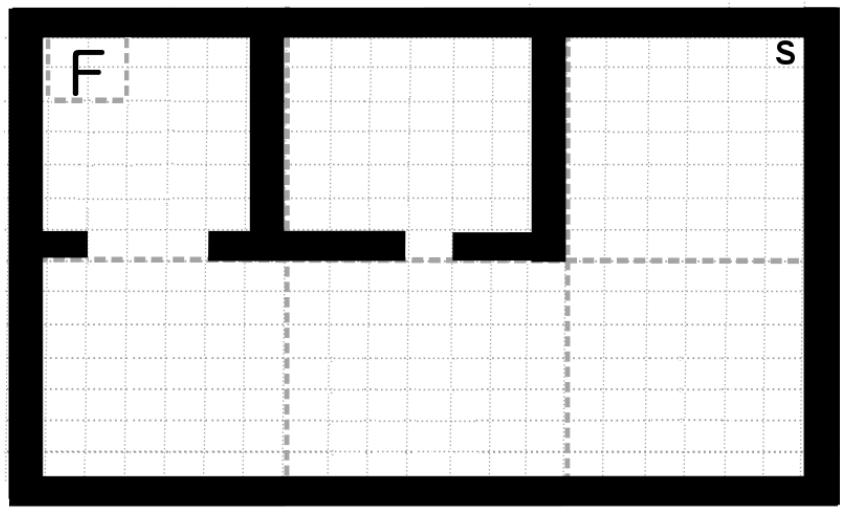
Kombinovaný funkčný aproximátor RBF-MLP

- riedke kódovanie s využitím lokálnych neurónov (Cetina 2008)



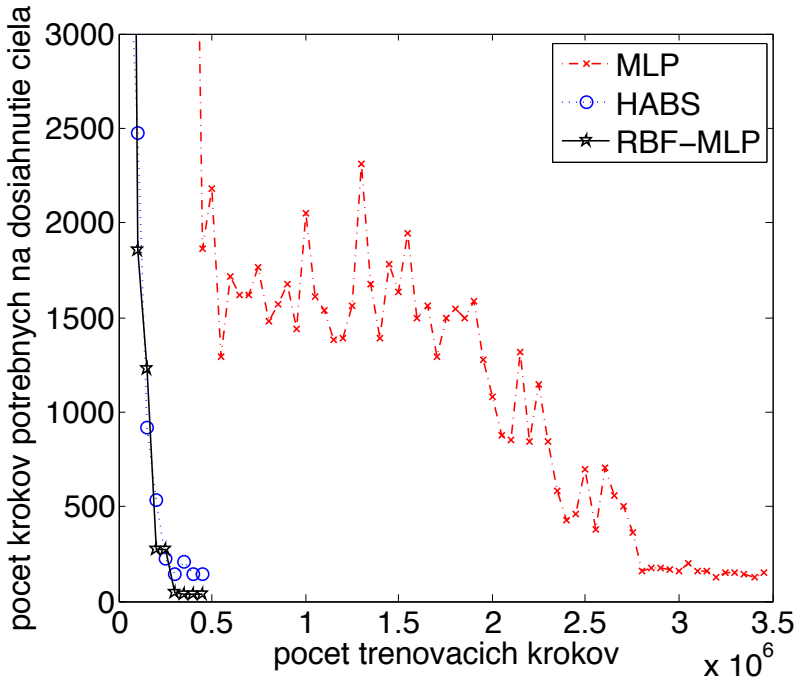
Prostredie

- Stav - pozícia agenta (x,y)
- Akcie - pohyb na 4 svetové strany
- Prechodová funkcia - deterministická, pri náraze do steny agent ostáva na mieste
- Odmenová funkcia - odmena 1 pri prechode do finálneho stavu, inak 0



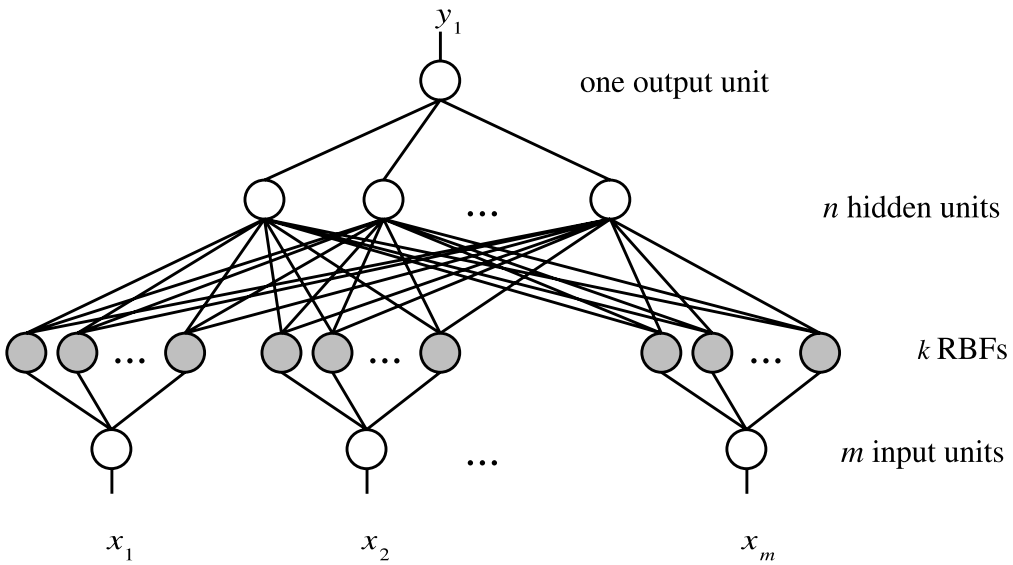
Výsledky

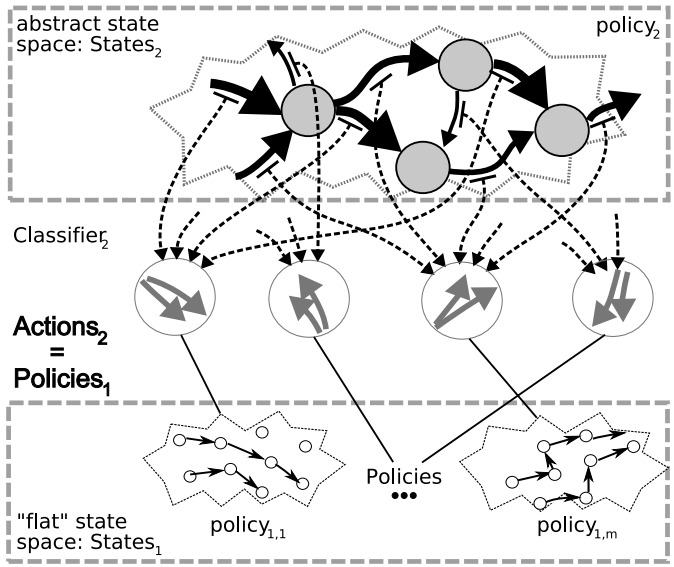
- MLP – 15 skrytých neurónov
- RBF-MLP – 15 skrytých neurónov, 6 RBF
- HABS – 4 podstratégie (2 skryté neuróny), hlavná stratégia (5 skrytých neurónov)



Porovnanie RBF-MLP a HABS

- veľmi podobný spôsob výberu akcie (lokálny + globálny výber)
- nižšia výpočtová náročnosť na jeden krok (HABS)





Záver

- problematika MDP, RL, HRL
- experimentálne porovnanie troch modelov – MLP, RBF-MLP a HABS
- hierarchické delenie priestoru znižuje dĺžku konvergencie
- časovo rozšírené akcie znižujú výpočtovú náročnosť jedného kroku

Ďakujem za pozornosť.

Viliam.Dillinger@fmph.uniba.sk