Aproximácia funkcie ohodnotení v algoritmoch Q-learning neurónovou sieťou

Michal CHOVANEC
Fakulta riadenia a informatiky

Október 2015

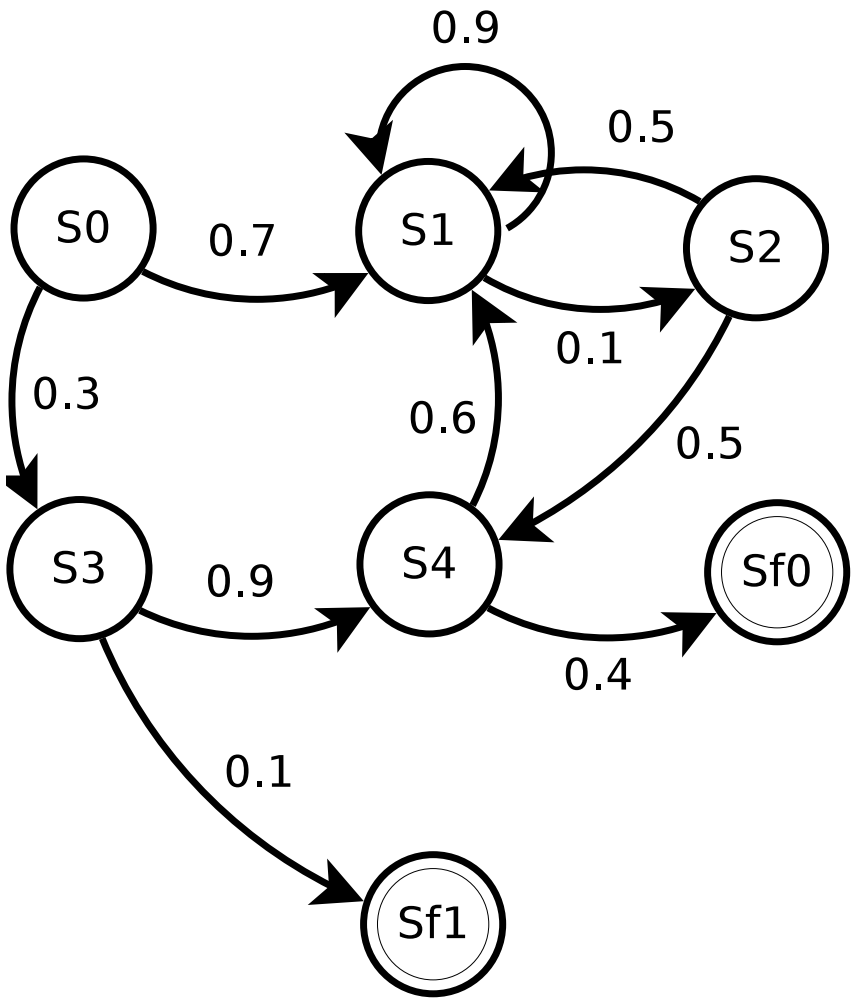
školiteľ : prof. Ing. Juraj Miček, PhD.
rok štúdia : 3.
nástup na štúdium : 1.9.2013

Úvod

Cieľom je nájsť optimálnu stratégiu - maximalizácia odmeny (účelovej funckie)

- Vopred nie je známa hodnota odmeny vykonanej akcie
- Vopred nie je známi ani stav do ktorého sa systém dostane
- Je možné určiť v akom stave sa systém nachádza
- Je presne daná množina akcií v každom stave
- Aspoň pre cieľový stav je daná výška odmeny

Q learning - prechod stavovým priestorom



Q learning - ohodnotenie Q(s, a)

$$Q_{n+1}(s,a) = R_{n+1}(s,a) + \gamma \max_{a'} Q_n(s'_{n+1}, a') \qquad (1)$$

Kde
$R_{n+1}(s,a)$ je získaná odmena (reward) za vykonanie akcie a v stave s v čase $n+1$

$\max_{a'} Q_n(s'_{n+1}, a')$ je výber akcie v stave s_{n+1} ktorá má najväčšiu odmenu

γ je podiel z maximálnej odmeny v stave s'_{n+1} pri vykonaní najlepšej možnej akcie v tomto stave

Q learning - ohodnotenie Q(s, a)

Varianty algoritmu
Filtrovanie v stochastickom prostredí

$$Q_{n+1}(s,a) = \alpha Q_n(s,a) + (1-\alpha)(R_{n+1}(s,a) + \gamma \max_{a'} Q_n(s'_{n+1}, a'))$$

SARSA algoritmus

$$Q_{n+1}(s,a) = \alpha Q_n(s,a) + (1-\alpha)(R_{n+1}(s,a) + \gamma Q_n(s'_{n+1}, a'))$$

kde $\alpha \in \langle 0, 1)$

Q learning - výber akcie

Boltzmanové rozdelanie

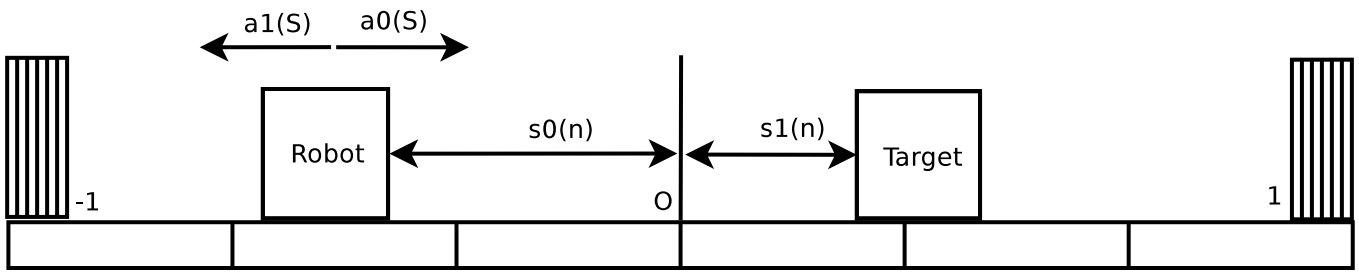
$$P(s|a_i) = \frac{k^{Q(s,a_i)}}{\sum_{j \in \mathbb{A}} k^{Q(s,a_j)}}$$

Kde $k \in \langle 0, \infty)$ a určuje správanie sa agenta, pre $Q(s,a) \in \langle -1, 1 \rangle$ možno pozorovať tieto druhy správania

- $k = 1$ prieskumník
- $k = \langle 2, 10 \rangle$
- $k >> 10$ pažravý (greedy) agent

Experiment

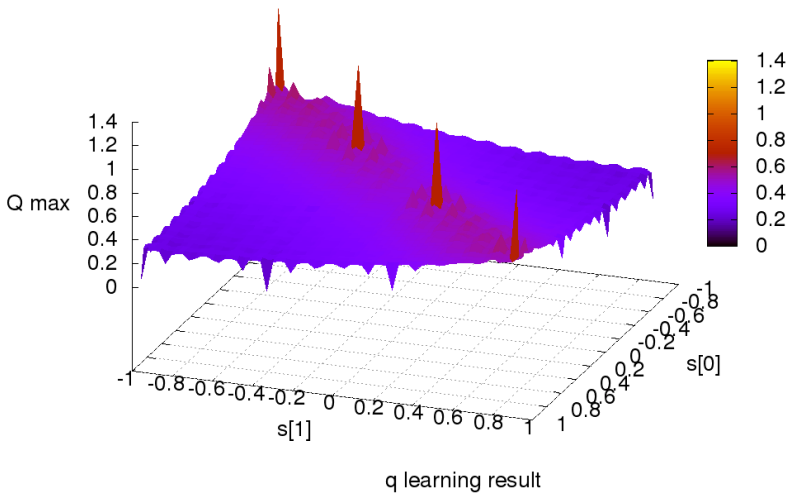
Cieľom je overiť aproximáciu $Q(s, a)$ dvoma rôznymi neurónovými sieťami pri rôznych veľkostiach k v malom stavovom priestore - Q(s, a) vieme spočítať presne.



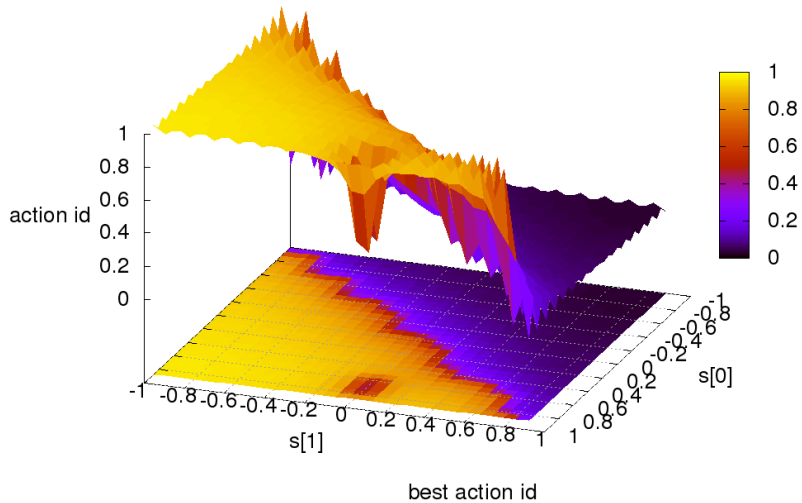
Experiment - parametre

```
iterations = 10000000

agent :

state_density = 1/8.0
alpha = 0.98
gamma = 0.7

neural network :

hidden layers = 2
neurons in hidden layers = 10
weight range = 4.0
neuron order = 7
init weight range = 0.1
eta = 0.001
```

Experiment, k = 1.1, optimálne riešenie

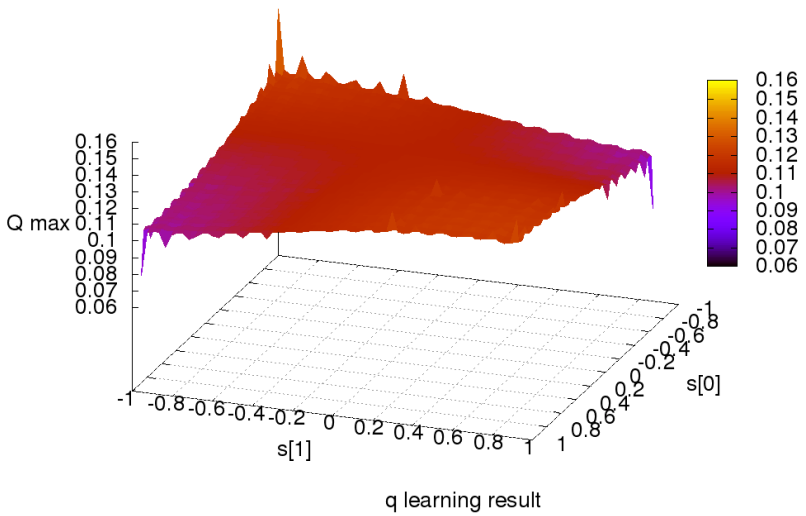


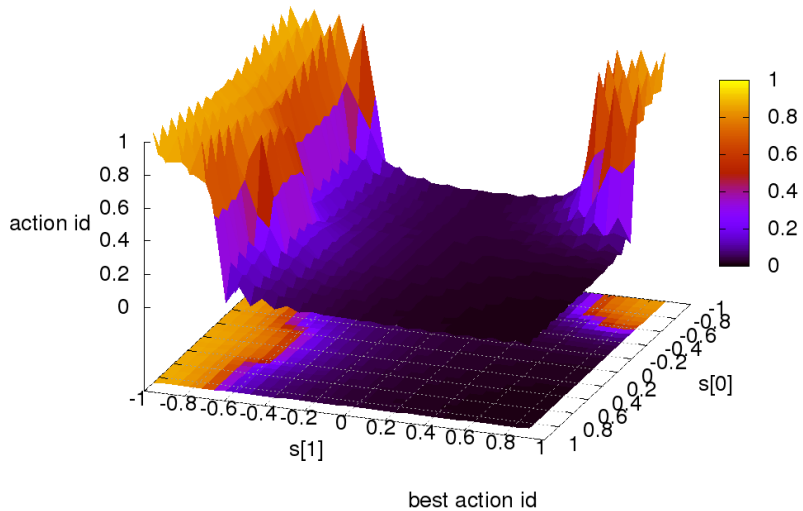
q learning result



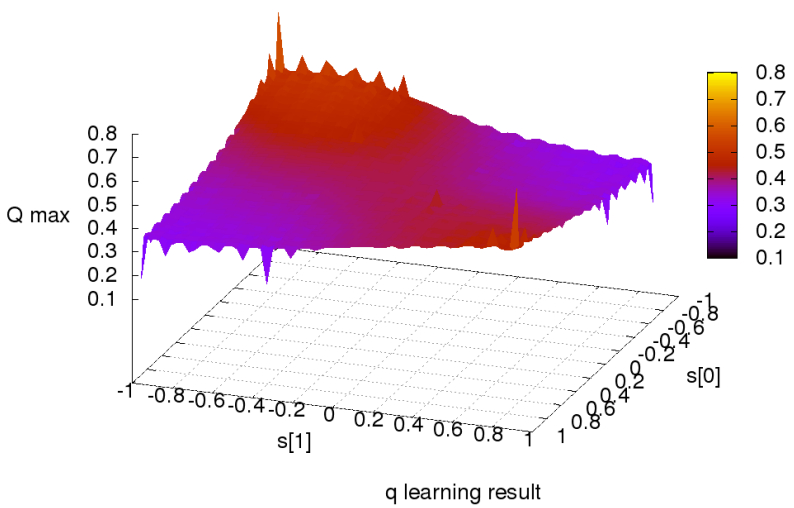
best action id

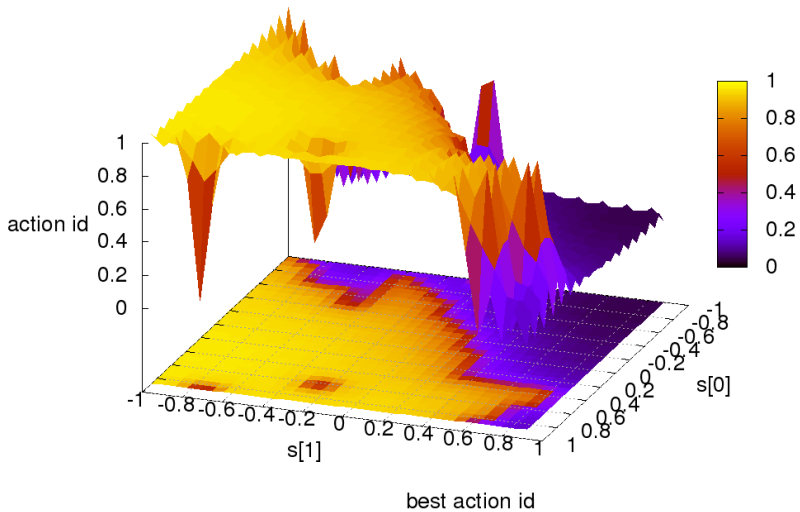
Experiment, k = 1.1, mcculloch pitts neurón



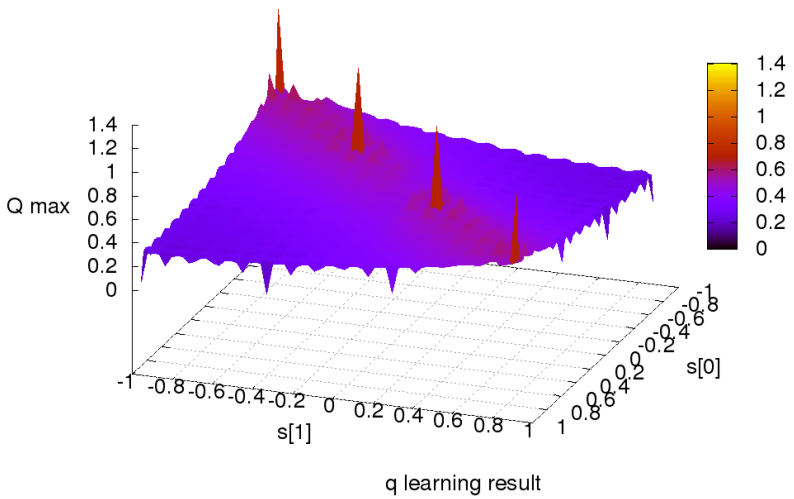


Experiment, k = 1.1, testovaný neurón

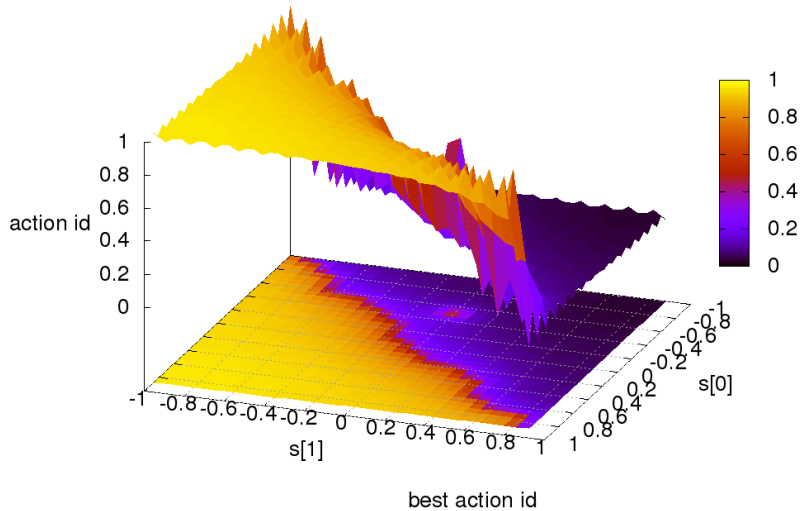




Experiment, k = 2.0, optimálne riešenie

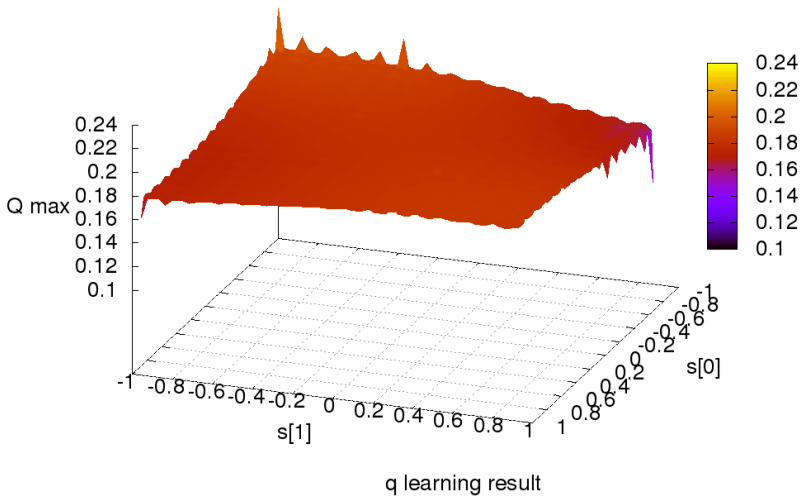


q learning result

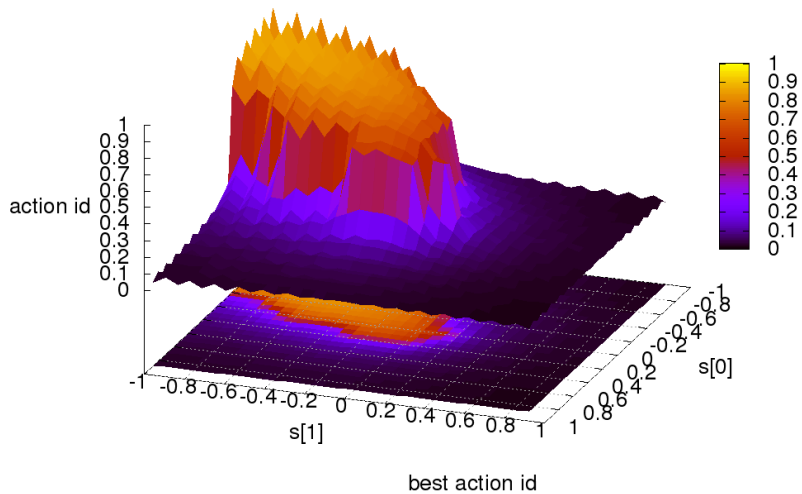


best action id

Experiment, k = 2.0, mcculloch pitts neurón

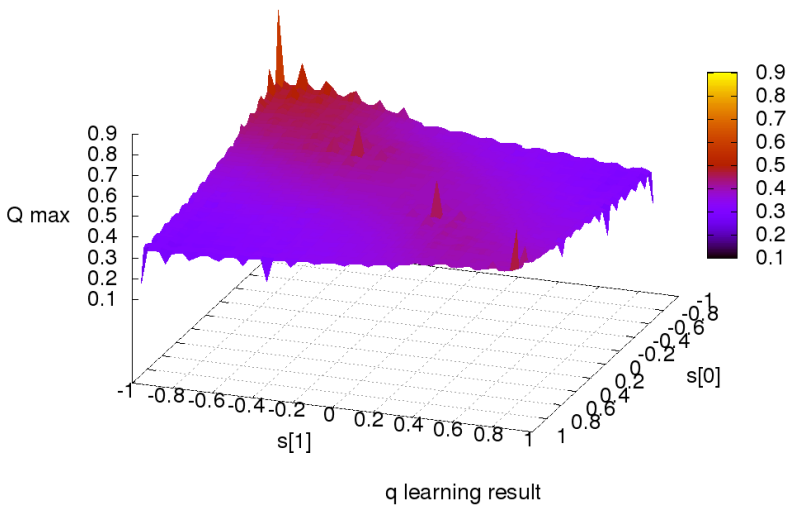


q learning result

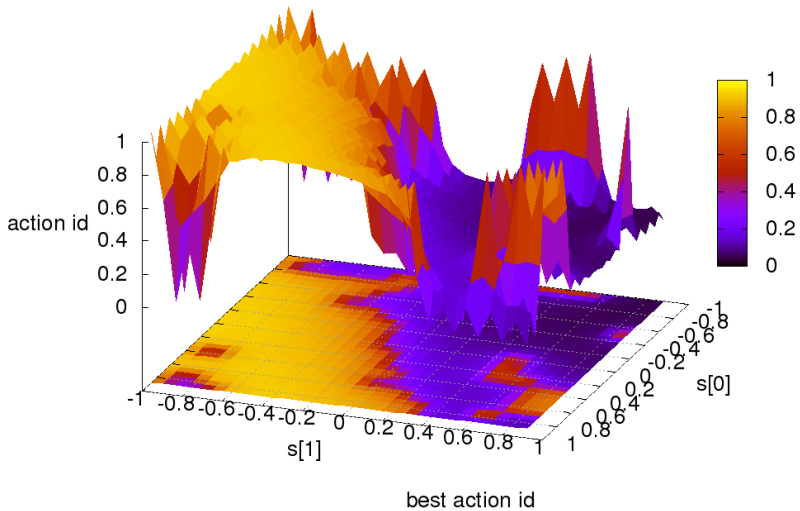


best action id

Experiment, k = 2.0, testovaný neurón

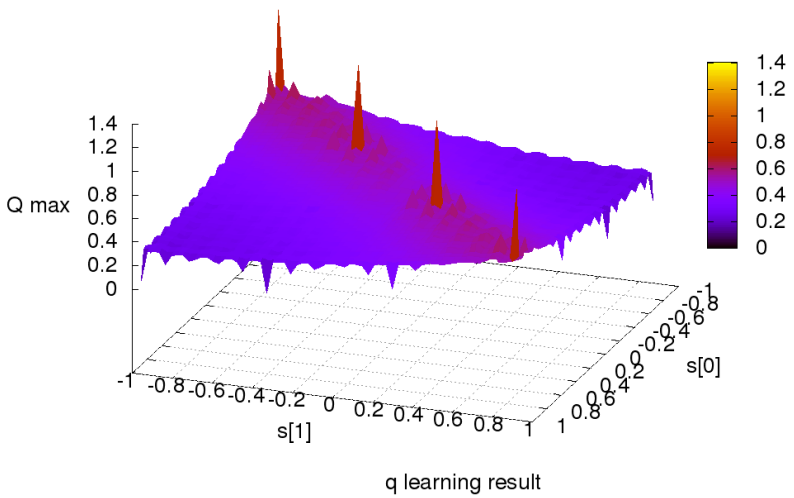


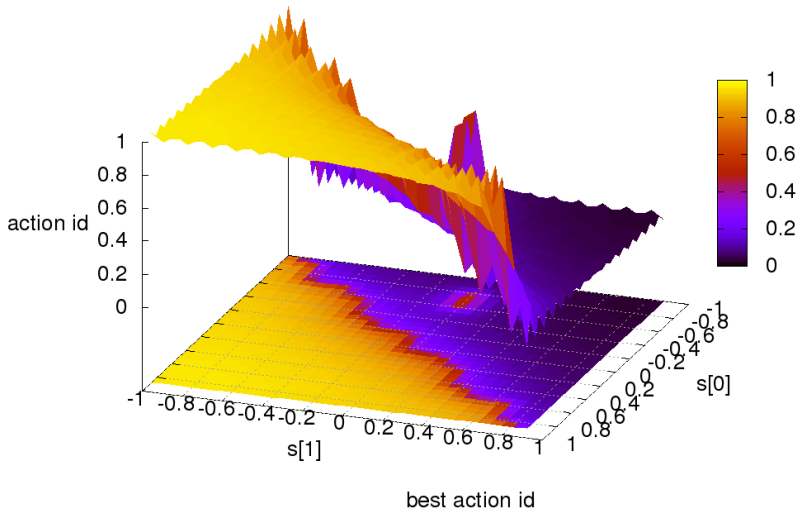
q learning result



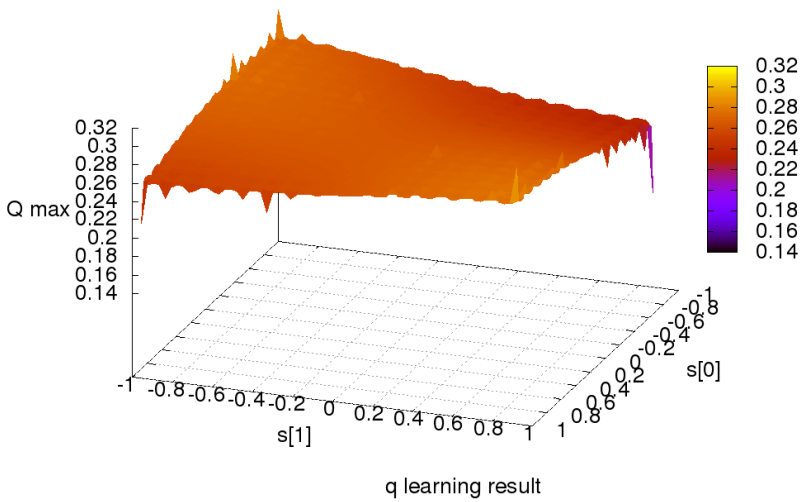
best action id

Experiment, k = 10.0, optimálne riešenie

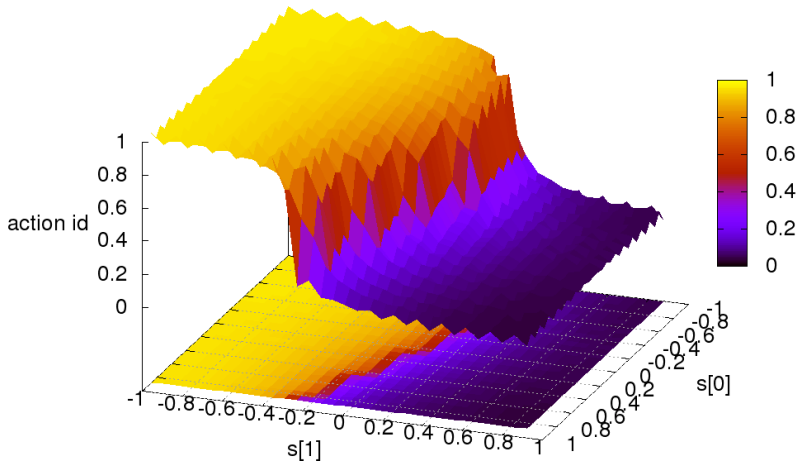




Experiment, k = 10.0, mcculloch pitts neurón

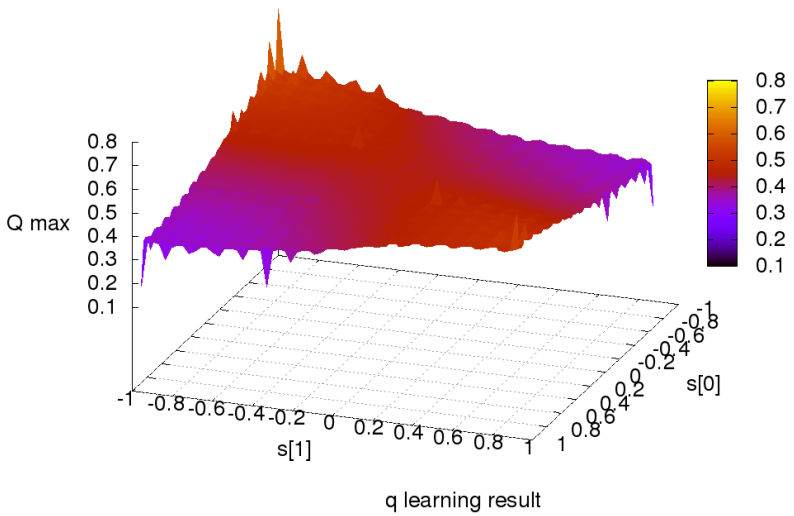


q learning result

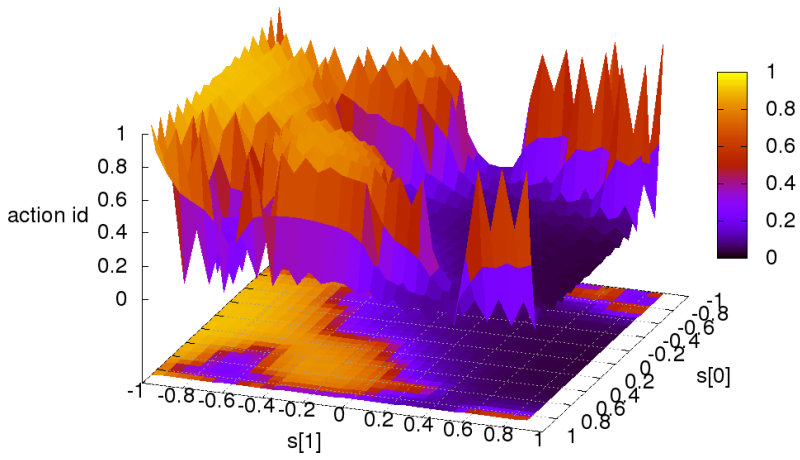


best action id

Experiment, k = 10.0, testovaný neurón



q learning result



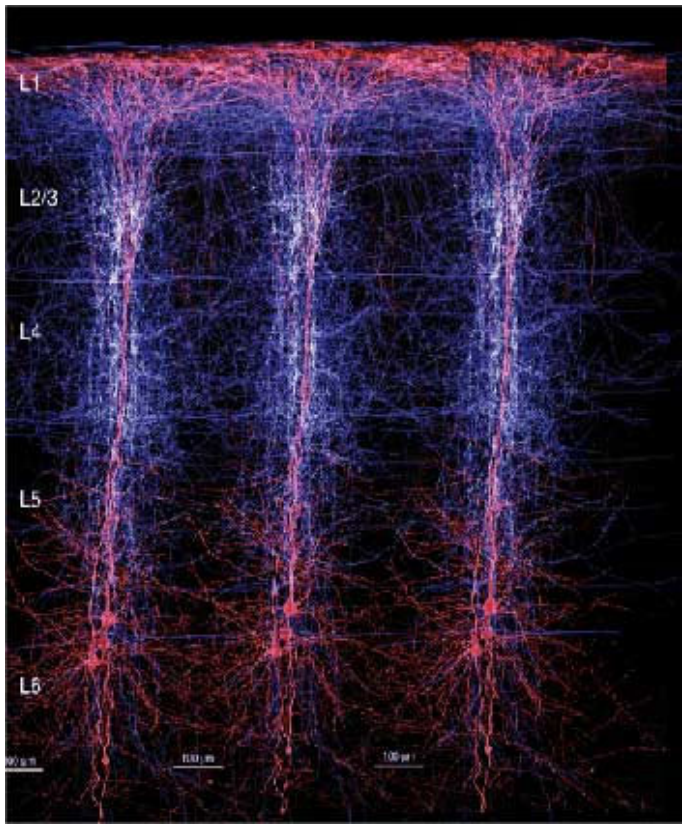
best action id

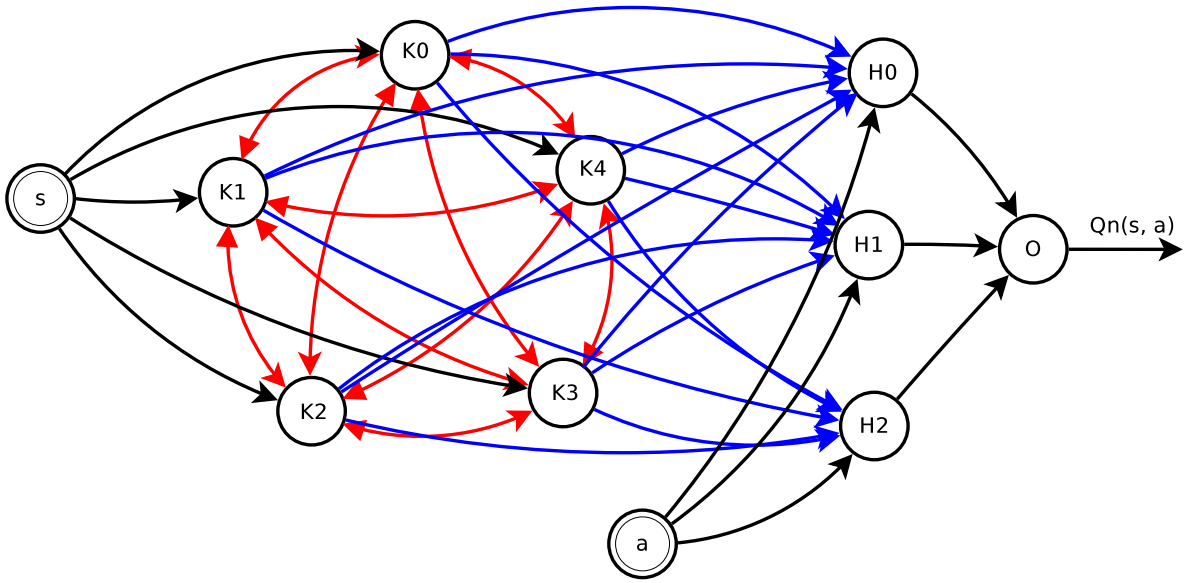
Experiment - zhrnutie

Neurón	k	chyba $\sum \lvert Q_o(s,a) - Q(s,a) \rvert$
McCulloch Pitts	1.1	95
Testing	1.1	37
McCulloch Pitts	2.0	110
Testing	2.0	50
McCulloch Pitts	10.0	79
Testing	10.0	69

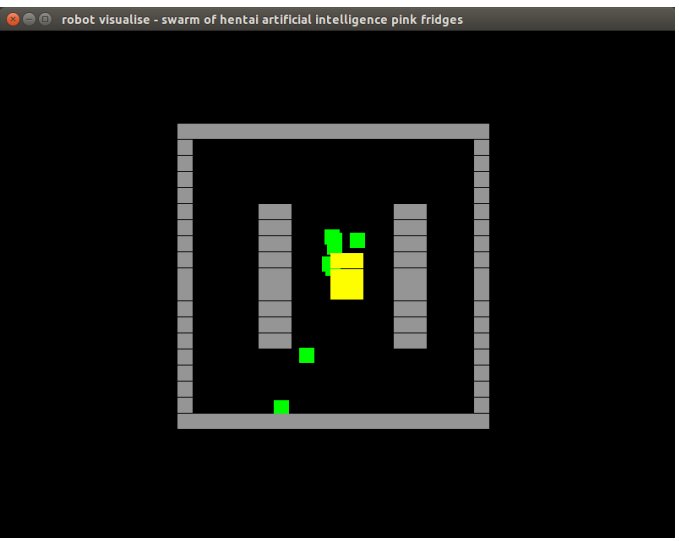
Prebiehajúci experiment

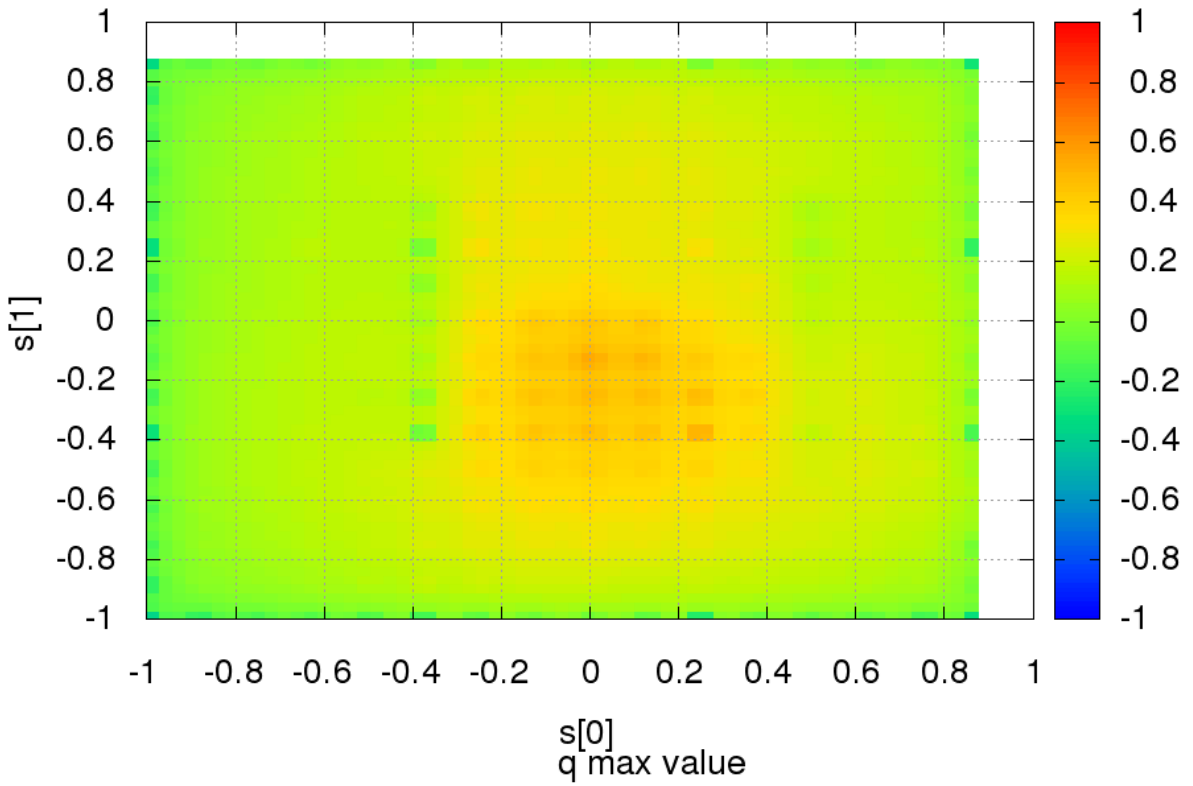
Overnie možnosti aproximácie Q(s, a) neurónovou sieťou, ktorej topológia vychádza z neokortexového stĺpca - stavebný kameň mozgovej kôry.



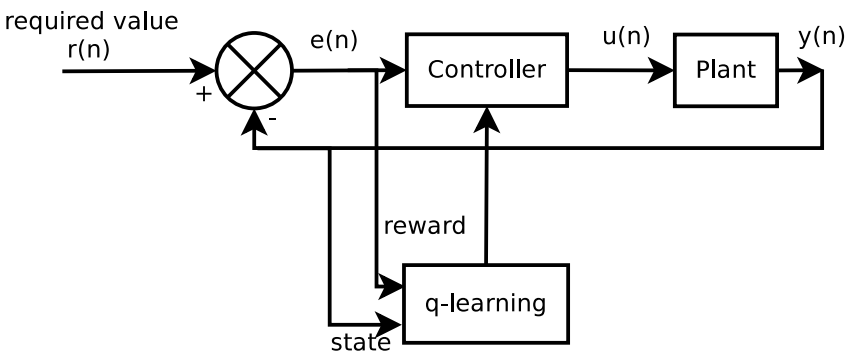


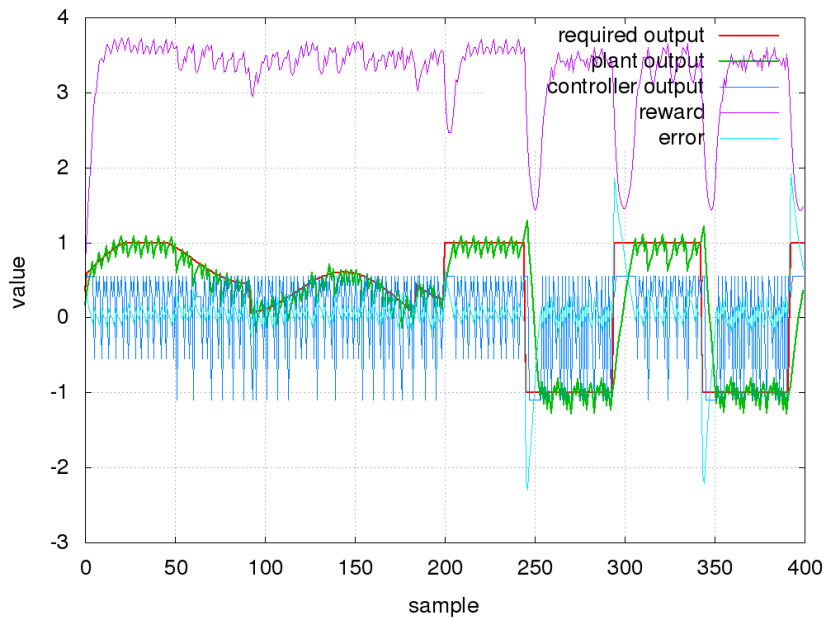
Prebiehajúci experiment - hľadanie cieľa na mape





Prebiehajúci experiment - učiaci sa regulačný systém





Ďakujem za pozornosť

michal.chovanec@yandex.com