

ŽILINSKÁ UNIVERZITA V ŽILINE
FAKULTA RIADENIA A INFORMATIKY

DIZERTAČNÁ PRÁCA

Študijný odbor: **Aplikovaná informatika**

Ing. Michal Chovanec

**Aproximácia funkcie ohodnotení v
algoritmoch Q-learning
neurónovou sieťou**

Vedúci: **prof. Ing. Juraj Miček, PhD**

Reg.č. xxx/2008

Máj 2016

Abstrakt

MICHAL CHOVANEC: *Aproximácia funkcie ohodnotení v algoritmoch Q-learning neurónovou sieťou* [Dizertačná práca]

Žilinská Univerzita v Žiline, Fakulta riadenia a informatiky, Katedra technickej kybernetiky.

Vedúci: prof. Ing. Juraj Miček, PhD

FRI ŽU v Žiline, 2016

Práca sa zaoberá aproximáciou funkcie ohodnotení konania agenta. V priestoroch s malým počtom stavov predstavuje vhodné riešenie tabuľka. Pre prípady veľkého počtu stavov je tabuľkové riešenie ťažko vypočítateľné. Je tak nutné použiť aproximáciu. Vhodným kandidátom je neurónová sieť. Tradičné riešenie doprednej siete je však nepoužiteľné z dôvodov nemožnosti takúto sieť učiť. V práci je preto venovaný priestor neurónovej sieti bázických funkcií ktorú už je možné na daný problém trénovať iteračnými metódami.

Abstract

MICHAL CHOVANEC: *Q-function approximation in Q-learning algorithms using neural network*

[Dissertation thesis]

University of Žilina, Faculty of Management Science and Informatics, Department of technical cybernetics.

Tutor: prof. Ing. Juraj Miček, PhD

FRI ŽU v Žiline, 2016

Práca sa zaoberá aproximáciou funkcie ohodnotení konania agenta. V preistoroch s malým počtom stavov predstavuje vhodné riešenie tabuľka. Pre prípady veľkého počtu stavov je tabuľkové riešenie ťažko vypočítateľné. Je tak nutné použiť aproximáciu. Vhodným kandidátom je neurónová sieť. Tradičné riešenie doprednej siete je však nepoužiteľné z dôvodov nemožnosti takúto sieť učiť. V práci je preto venovaný priestor neurónovej sieti bazických funkcií ktorú už je možné na daný problém trénovať iteráčnými metódami.

Prehlásenie

Prehlasujem, že som túto prácu napísal samostatne a že som uviedol všetky použité pramene a literatúru, z ktorých som čerpal.

V Žiline, dňa 15.5.2012

Meno Priezvisko

Obsah

1	Úvod	4
1.1	Markovove rozhodovacie procesy	4
2	Učiace sa systémy na báze odmeňovania	6
2.1	Adaptívny systém	6
2.2	Princípy učenia s odmeňovaním	7
2.3	Ohodnocovanie vykonaných akcií	9
2.4	Jednostavový systém	12
2.4.1	Výsledky experimentu	13
3	Q-larning algoritmus	21
3.1	Definícia algoritmu	21
3.2	Výber akcie	25
3.3	Problémy výpočtu $Q(s,a)$	26
3.4	Tabuľka	27
3.5	Dopredná neurónová sieť	28
3.6	Kohonenová neurónová sieť	30
3.7	Neurónová sieť bazických funkcií	33
3.7.1	Určenie parametrov α	36
3.7.2	Určenie parametrov β	37
3.7.3	Určenie váhových parametrov w	37
3.7.4	Hybridný variant	38

	3
4 Experimentálna časť	39
4.1 Ciele epperimentu	39
4.2 Návrh experimentu	40
4.3 Výsledky experimentu	43
Literatúra	52

Kapitola 1

Úvod

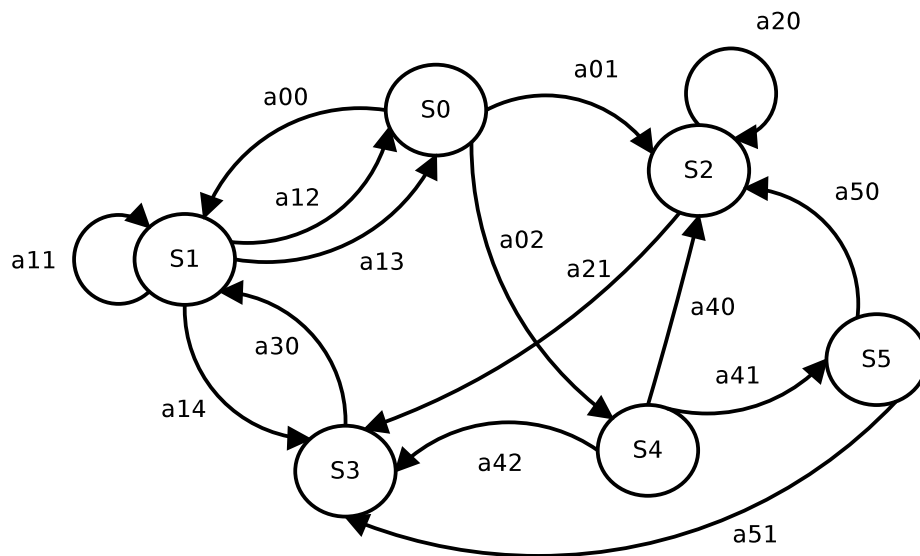
Strojové učenie predstavuje ďalší logický krok v počítačovej vede. Programovanie založené na pevne danom správaní prestáva pre mnohé úlohy stačiť. Na softvérové produkty sú kladené čoraz väčšie nároky, a v dohľadnej dobe sa dá očakávať dosiahnutie hranice tradičného prístupu. Dôvodom je najmä veľmi rozsiahly stavový priestor úloh. Možným východiskom sú adaptívne a učiace sa systémy.

1.1 Markovove rozhodovacie procesy

Množstvo úloh je možné popísať množinou stavov \mathbb{S} , pre každý stav je definovaná množina akcií $\mathbb{A}_{\mathbb{S}}$. Pre každý stav a každú akciu, ktorú je v ňom možné vykonať existuje pravdepodobnostná funkcia prechodu do ďalšieho stavu $P(s, s')$ a za uskutočnené prechody je daná funkcia odmien $R(s, s')$. Formálne je teda Markovov proces päťica $(\mathbb{S}, \mathbb{A}_{\mathbb{S}}, P(s, s'), R(s, s'), \gamma)$, kde $\gamma \in (0, 1)$ je faktor zabúdania, a volí sa ako parameter procesu. Jeho význam bude vysvetlený v ďalšej kapitole. Príklad Markovovho rozhodovacieho procesu je na obrázku 1.1. Prechody v grafe znamenajú jednotlivé akcie.

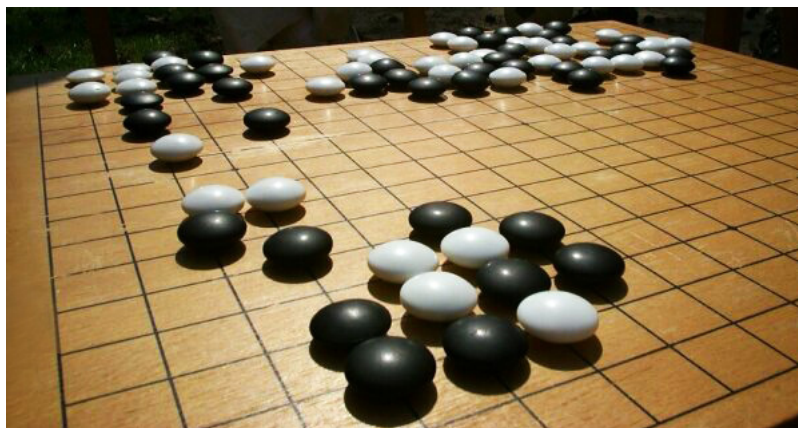
V priestore stavov \mathbb{S} existuje výkonná jednotka - agent. Je to modul ktorý na základe vstupnej informácie - stave vyberie jednu z akcií. To s akou pravdepodobnosťou sa daná akcia vykoná je dané prostredím (popísané Markovove rozhodovacím procesom).

Dôležitým východiskom je, že $P(s, s')$ nie je agentovi známa. Agent teda nemôže vedieť kam sa vykonaním vybranej akcie dostane. Tento fakt značne komplikuje prehľadávanie



Obr. 1.1: Markovov rozhodovací proces

priestoru. Ďalším typickým rysom je, že odmeňovacia funkcia $R(s, s')$, ktorá je zadaná, nadobúda pre väčšinu prechodov z s do s' hodnotu 0. Je to dôsledok prirodzeného spôsobu riešenia problémov delením na čiastkové kroky a nie je okamžite známa veľkosť úžitku po vykonaní čiastkového kroku. Často je táto skutočnosť známa len pre pár vybraných prechodov. Dobrým príkladom tejto situácie je hra Go.



Obr. 1.2: Hra Go

Položením kameňa na Goban nie okamžite známe ako to ovplyvní priebeh hry. Je bežné, že sa to prejaví až po mnohých krokoch.

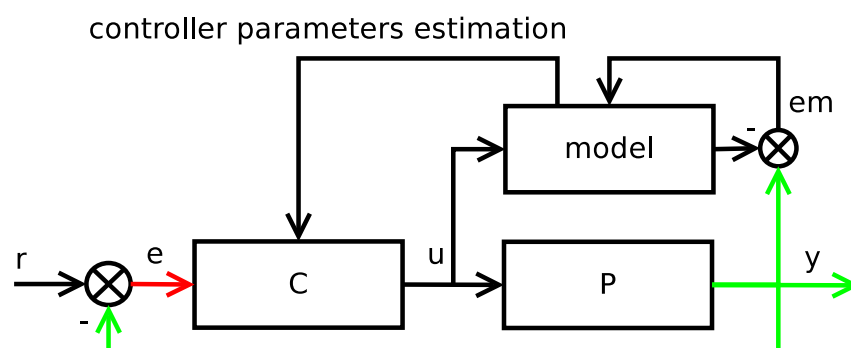
Kapitola 2

Učiace sa systémy na báze odmeňovania

V tejto kapitole budú stručne predstavené učiace sa systémy založené na odmeňovaní. Je nevyhnutné, aspoň okrajovo spomenúť adaptívne systémy riadenia, aby bolo možné pochopiť význam a rozdiel oproti predmetným systémom. Kapitola si kladie za cieľ ukázať princípy, matematické detaily budú rozobraté v ďalšej kapitole.

2.1 Adaptívny systém

V teórii riadenia sa narába s pojmami spätná väzba, žiadaná hodnota, chybová veličina, riadená sústava, regulátor a model. Obrázok 2.1 zobrazuje prepojenie základných blokov typického adaptívneho systému.



Obr. 2.1: Adaptívny riadiaci systém

Pre ďalšie účely sa bude hovoriť o systémoch pracujúcich v diskretnom čase n s krokom

1. Vstupom je žiadaná hodnota $r(n)$, výstupom je $y(n)$. Z toho nevyhnutne vyplýva definícia chybovej veličiny ako $e(n) = r(n) - y(n)$. Kvalitu regulátora je možné definovať rôznymi spôsobmi, najbežnejší je

$$K(c) = \sum_{n=1}^c e(n)^2 \quad (2.1)$$

Cieľom je minimalizovať $K(c)$. Aby existoval súčet tohto radu musí platiť $\lim_{n \rightarrow \infty} e(n)^2 = 0$.

Regulátor $C(n)$ je blok, tvorený obvykle diferenčnou rovnicou s volenými parametrami do ktorého vstupuje chyba. Veličina $u(n)$ je výstup regulátora a nazýva sa akčná veličina, predstavuje vstup do riadenej sústavy $C(n)$. Akčná veličina ďalej vstupuje do modelu systému ktorého parametre sa menia podľa chyby modelu $e_m(n) = y(n) - y_m(n)$, kde $y_m(n)$ je výstup modelu. Cieľom je minimalizovať $L(c) = \sum_{n=1}^c e_m(n)^2$. Po dosiahnutí minima sa parametre modelu rovnajú parametrom riadenej sústavy, a je tak možné urobiť syntézu regulátora.

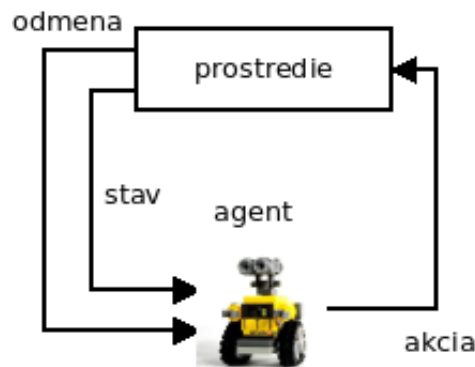
2.2 Princípy učenia s odmeňovaním

Pre všetky n musí existovať požadovaná hodnota $r(n)$ a k nej prislúchajúca chyba $e(n)$. Čo v prípade systémov kedy nie je pre všetky n definované správanie? Príkladom môže byť rozhodovanie robota, ktorý má splniť cieľ pozostávajúci z niekoľkých elementárnych úkonov, ale postupnosť týchto elementárnych úkonov nie je známa - nie je teda definované $r(n)$ pre každé n .

Riešením tohto problému je zavedenie systému odmeňovania agenta (robota) 2.2.

Táto metóda sa nazýva **reinforcement learning**. V súčasnosti predstavuje najmocnejší nástroj strojového učenia v oblasti umelej inteligencie. Umožňuje riešiť veľmi komplexné problémy a uplatnenie nachádza nie len v počítačových hrách, ale napr. aj v robotike. Jej princíp je možné zhrnúť do týchto bodov :

1. Zistenie stavu
2. Výber akcie



Obr. 2.2: Učenie s odmeňovaním

3. Vykonanie akcie
4. Prechod do ďalšieho stavu
5. Získanie odmeny alebo trestu
6. Učenie sa zo získanej skúsenosti

Zistenie stavu je umožnené vnímaním prostredia, napr. pomocou senzorov. V najjednoduchšom prípade si je možné ako stav predstaviť polohu. Vo všeobecnosti, to ale nemusia byť ani priamo dáta so vstupov, ale len nejaké príznaky - robot je na križovatke $p_0 = 0.9$, robot je v uličke $p_1 = 0.3$. Sada príznakov umožňuje zredukovať dimenziu stavového priestoru (tá musí byť pre ďalej rozoberané algoritmy konečná, vyplynie to z prezentovaných rovníc).

Výber akcie je uskutočnený na základe ohodnotení akcií získaných v minulosti. Tieto ohodnotenia sú samozrejme závislé od stavu v ktorom sa systém nachádza - tá ista množina akcií má iné ohodnotenie keď je robot na križovatke a iné keď je v uličke. Vo fáze prieskumu prostredia môže agent vyberať tie akcie ktoré boli v minulosti málo často vykonané, alebo môže vyberať náhodne. Definovaním veľkosti zmeny ohodnotenia akcie môže uprednostniť tie akcie, ktorých ohodnotenia vykazujú veľkú zmenu - pretože je pravdepodobné že sa tým dostane do nepreskúmaných stavov. Vo fáze kedy agent už nepreskúma prostredie, môže vyberať len najlepšie ohodnotenú akciu. Prípadne môžu existovať v prostredí s viacerými agentami prieskumníci a vykonávatelia. Navzájom si môžu vymieňať skúsenosti.

Vykonanie akcie V deterministickom prostredí sa vybraná akcia vykoná vždy rovnako

(opäť však závisí na stave agenta). V nedeterministickom prostredí sa vykonanie sa akcie riadi nejakou pravdepodobnostnou funkciou - napr. robotovi môže prešmyknúť koleso.

Prechod do ďalšieho stavu je obvyklým dôsledkom vykonania akcie. Samozrejme, že v danom stave môže existovať aj taká akcia ktorá nespôsobí zmenu stavu - napr. spomínané prešmyknutie kolesa - robot sa nepohne.

Získanie odmeny alebo trestu po vykonaní akcie, existuje odmeňovací mechanizmus zadaný tvorcom systému. Preto sa nedá hovoriť o umelej inteligencii - správanie sa agenta je v deterministickom prostredí plne určené týmto mechanizmom - len nie je známe. Obvykle platí, že kladné hodnoty dosahuje odmena, záporné trest. Nulová hodnota symbolizuje neznalosť zadavateľa, a nevie teda určiť či dané konanie bolo dobré alebo zlé. Nesmierne dôležitý je fakt, že pre veľkú väčšinu rozhodnutí je výsledkom odmeny práve nula. Príkladom môže byť japonská hra Go - pre úspešné vedenie partie nestačí ohodnotiť len aktuálny stav na gobane, je potrebné uvažovať viac ťahov dopredu - úspešnosť ťahu sa často ukáže až po vykonaní množstva ďalších ťahov, kedy je odmena už nenulová.

Práve dominancia nulových odmien je unikátom pre reinforcement learning algoritmy. Od zadavateľa sa vyžaduje minimum znalostí, je dokonca možné ukázať, že stačí odmena len v cieľovom stave (i keď proces učenia bude v tomto prípade trvať veľmi dlho).

2.3 Ohodnocovanie vykonaných akcií

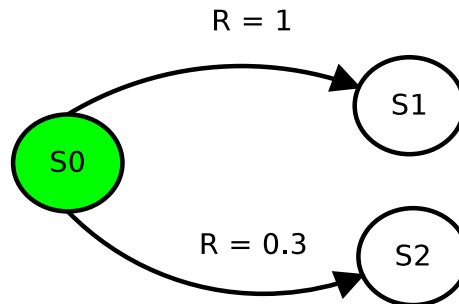
Odmena je získaná z prostredia po vykonaní akcie. Ohodnotenie akcie v danom stave je tvorené predoslymi skúsenosťami z vykonania predošlých akcií a získania odmien. Podstatou učenia je teda ohodnotenie vykonaných akcií v danom stave, aby bolo možné v každom stave rozhodnúť ktorá akcia je najlepšia - vyberá sa teda postupnosť akcií π pre ktorú je funkcia

$$\Lambda(\pi) = \sum_{n=0}^{\infty} \gamma^n R_{\pi(n)}(s(n), s(n-1)) \quad (2.2)$$

maximálna. Kde $\gamma \in \langle 0, 1 \rangle$ je koeficient zabúdania, $R_{\pi(n)}(s(n), s(n-1))$ je odmeňovacia funkcia po prechode zo stavu $s(n-1)$ do stavu $s(n)$ vykonaním $\pi(n)$.

Hľadanie mechanizmu nájdenia maxima $\Lambda(\pi)$ bude vysvetlené na nasledujúcich príkladoch.

Pre ilustráciu bude postupnosť prechodov zapísaná priamo ako parameter funkcie Λ , napr : $\Lambda(\{S_0, S_7, S_5\})$ predstavuje prechody z S_0 do S_7 a z S_7 do S_5 . Ďalej sa predpokladá že $\gamma = 1$. Je daný systém s troma stavmi, a dvoma prechodmi 2.3. Pre ilustráciu je systém tak jednoduchý, že každá akcia má známu odmenu.



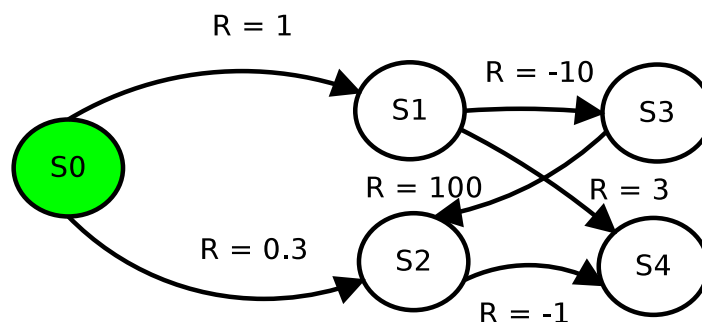
Obr. 2.3: Ohodnocovanie akcií v trojstavovom systéme

V tomto systéme je ohodnotenie $\Lambda(S_a, S_b)$ naozaj triviálne

1. $\Lambda(\{S_0, S_1\}) = 1$
2. $\Lambda(\{S_0, S_2\}) = 0.3$

Najlepšia cesta je potom $\{S_0, S_1\}$.

V prípade systému s viacerými stavmi 2.4, ale opäť známymi ohodnoteniami v každom prechode bude situácia nasledovná.



Obr. 2.4: Ohodnocovanie vo viacstavovom systéme

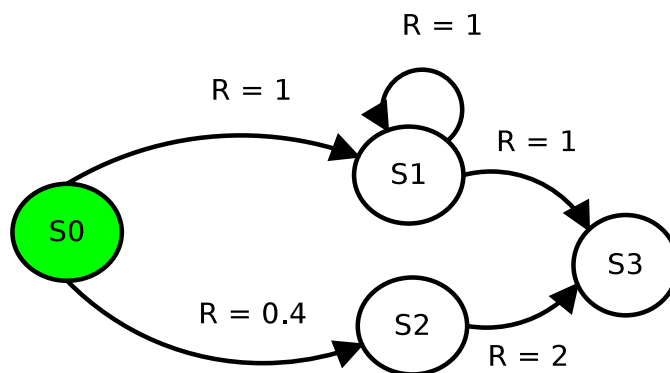
Ohodnotenie ciest :

1. $\Lambda(\{S_0, S_1, S_3\}) = 1 + (-10) = -9$

2. $\Lambda(\{S_0, S_1, S_4\}) = 1 + 3 = 4$
3. $\Lambda(\{S_0, S_2, S_4\}) = 0.3 + () - 1 = -0.7$
4. $\Lambda(\{S_0, S_1, S_3, S_2, S_4\}) = 1 + (-10) + 100 + (-1) = 90$
5. ...

Jednoduchých sčítaním ohodnotení rôznych ciest je tak možné nájsť optimálnu postupnosť akcií.

Ak sa v systéme nachádza cyklus 2.5 (na obrázku je triviálny prípad) ohodnotenie bude divergovať. Agent bude mať možnosť vykonávať akciu ktorá neustále pripočítava kladnú odmenu, nevyhnutne tak vyberie túto stratégiu, pretože tak získa nekonečne veľkú odmenu.



Obr. 2.5: Ohodnocovanie s cyklom

Ohodnotenie niektorých ciest

1. $\Lambda(\{S_0, S_2, S_3\}) = 0.4 + 2 = 2.4$
2. $\Lambda(\{S_0, S_1, S_3\}) = 1 + 1 = 2$
3. $\Lambda(\{S_0, S_1, S_1, S_1\}) = 1 + 1 + 1 = 3$
4. $\Lambda(\{S_0, S_1, S_1, S_1, S_1\}) = 1 + 1 + 1 + 1 = 4$
5. $\Lambda(\{S_0, S_1, S_1, S_1, S_1, S_1, \dots\}) = 1 + 1 + 1 + 1 + 1 + \dots + 1 = \infty !!!!$

Riešením je práve zavedenie faktora zabúdania γ . Ilustračne bola zvolená $\gamma = 0.9$. Agent tak síce urobí niekoľko cyklov S_1, S_1 , po čase ale hodnota odmeny konverguje a minulé príspevky dávajú čoraz menší a menší úžitok a nevyhnutne tak po niekoľkých cykloch zmení rozhodnutie a prejde z S_1 do S_3 .

Ohodnotenie niektorých ciest

1. $\Lambda(\{S_0, S_2, S_3\}) = 2 + 0.9 * 0.4 = 2.36$
2. $\Lambda(\{S_0, S_1, S_3\}) = 1 + 0.9 * 1 = 1.9$
3. $\Lambda(\{S_0, S_1, S_1, S_1\}) = 1 + 0.9 * (1 + 0.9 * 1) = 2.71$
4. $\Lambda(\{S_0, S_1, S_1, S_1, S_1\}) = 1 + 0.9 * (1 + 0.9 * (1 + 0.9 * 1)) = 3.439$
5. $\Lambda(\{S_0, S_1, S_1, S_1, S_1, S_1, \dots\}) = 10 < \text{---}$
6. $\Lambda(\{S_0, S_1, S_1, S_1, S_1, S_1, \dots, S_3\}) = 11 < \text{---}$

Cyklovanie v S_1 teda prinesie celkovú odmenu 10, ale pri prechode do S_3 je celková odmena 11.

Formálny prepis uvedených úvah vedie na Q-learning algoritmus, ktorý zohľadňuje Bellmanov princíp optimality.

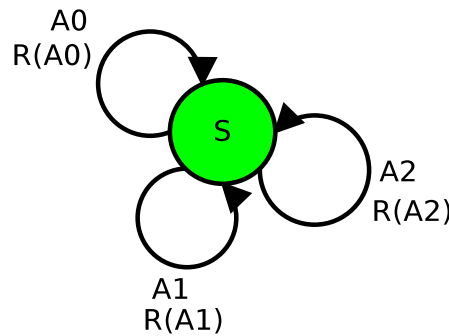
2.4 Jednostavový systém

Prv než bude uvedené úplne znenie Q-learning algoritmu je vhodné rozobrať ešte jeden príklad, a to systém s jedným stavom 2.6. Tento problém vznikol z nasledujúcej úvahy : je daný robot ktorý má jeden senzor vzdialenosti od cieľa (nevie teda určiť smer (napr. senzor intenzity osvetlenia)) a možnosti pohybu o elementárny krok vpred, vľavo, vpravo, vzad. Ako zvoliť postupnosť akcií aby robot došiel do cieľa - minimalizoval vzdialenosť? Navyše sa predpokladá, že senzor je nekvalitný a záludný zároveň : poskytuje informáciu len o tom či sa situácia zlepšila alebo zhoršila (oproti predošlému meraniu), a s určitou pravdepodobnosťou generuje náhodnú hodnotu - náhodné zmení výsledok (šum).

To vedie na zostavenie odmeňovacej funkcie ako

$$R(n) = \begin{cases} k & \text{ak } d(n) - d(n-1) < 0 \\ -k & \text{inak} \end{cases} \quad (2.3)$$

kde $d(n)$ je zmeraná vzdialenosť a na $d(n) - d(n-1) < 0$ sa pozerá ako uzavretú časť - vstupon do algoritmu teda nie je samotné $d(n)$ ale až $R(n)$. Pre k platí : $k = 1$ ak $rnd(0,1) > p$,inak -1 , kde p je pravdepodobnosť zmeny nameranej hodnoty $p \in \langle 0,1 \rangle$ a $rnd(0,1)$ generuje náhodné číslo z intervalu $\langle 0,1 \rangle$.



Obr. 2.6: Jednostavový systém s troma akciami

Kľúčový je výber akcie $a(n)$ z konečnej množiny akcií \mathbb{A} , vychádza sa z predstavy : ak má vykonaná akcia dobré ohodnotenie, vykoná sa znova, ak má zlé ohodnotenie, vyberie sa iná, náhodná akcia. Formálne zapísané

$$a(n) = \begin{cases} a(n-1) & \text{ak } Q_{n-1}(a(n-1)) > 0 \\ random() & \text{inak} \end{cases} \quad (2.4)$$

Samotné Q predstavujúce ohodnotenie sa potom vypočíta ako

$$Q_n(A(n)) = R(n) + \gamma \max_{a'(n-1) \in \mathbb{A}} Q_{n-1}(a'(n-1)) \quad (2.5)$$

Algoritmus je teda veľmi jednoduchý a možno ho použiť na plánovanie pohybu robota. Robotovi tak nie vnútené správanie programátorom, ale sám sa naučí ktoré akcie vyberať.

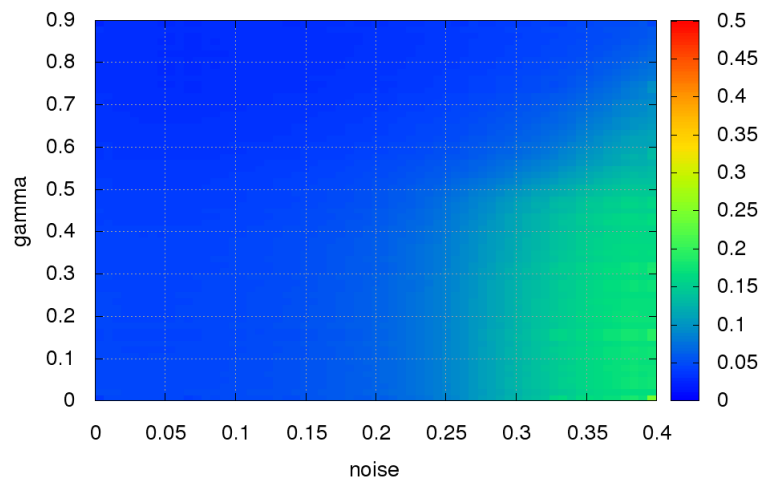
2.4.1 Výsledky experimentu

Overenie prebehlo s 32 virtuálnymi robotmi. Cieľom bolo naháňať pohyblivý cieľ, ktorý sa pohyboval po kružnici. Počiatočné polohy robotov boli náhodné. Celý experiment prebiehal

v dvojrozmernom priestore obmedzenom na rozsah polohy $\langle -1, 1 \rangle$. Metrika vzdialenosti bola Euklidovská.

Po 25000 iteráciach sa počítala priemerná chyba z 32 robotov. Vyšetrovala sa závislosť tejto chyby od zmeny parametrov γ a p (na grafe ako gamma a noise). Pre vybrané prípady boli exportované aj dráhy prvých 8 robotov.

Prostredie sa tak postupne mení z plne deterministického bez šumu až po hodnotu 0.4, čo predstavuje 40% pravdepodobnosť zmeny hodnoty senzora. Graf závislosti priemernej chyby od parametrov γ a p je na obrázku 2.7.

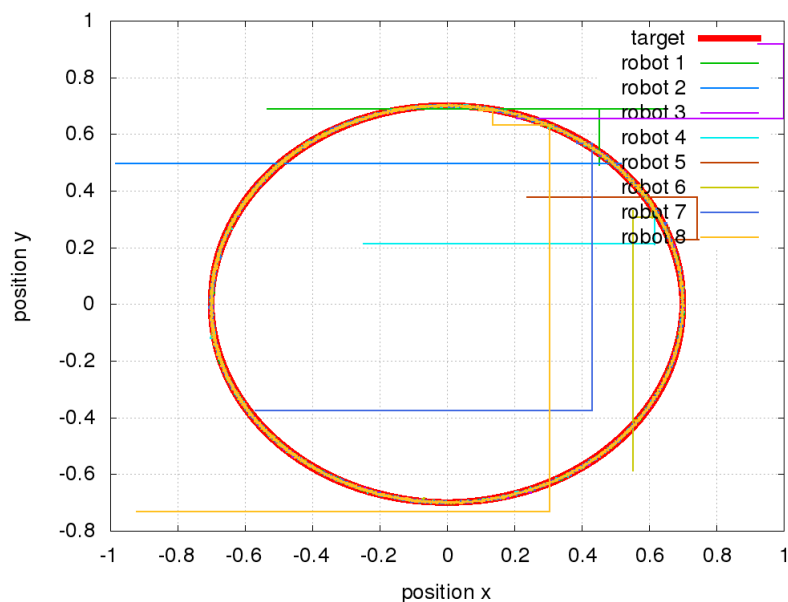


Obr. 2.7: Priemerná chyba 32 robotov po 25000 iteráciach

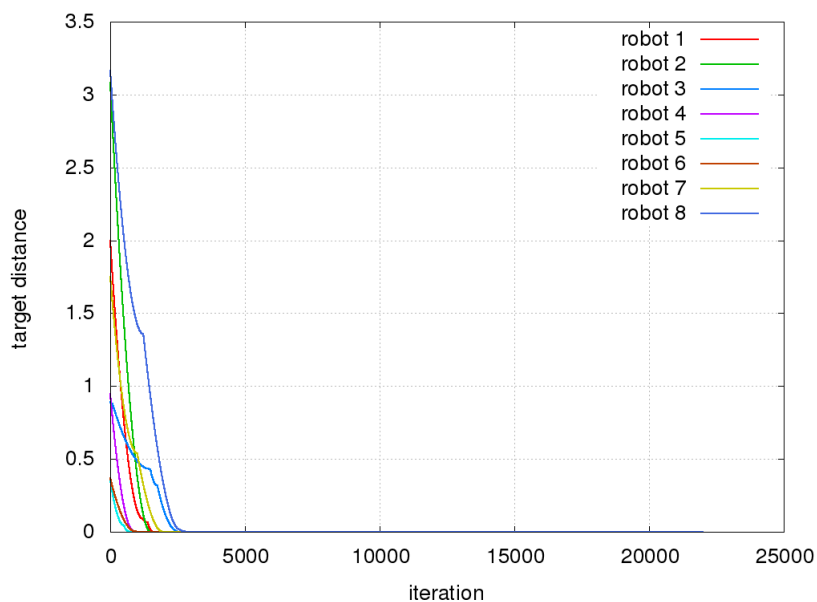
Dráhy robotov pre niektoré zvolené parametre sú uvedené v tabuľke 2.1

γ	p	graf dráhy	graf vývoja chyby
0.7	0.0	2.8	2.9
0.7	0.3	2.10	2.11
0.0	0.4	2.12	2.13
0.7	0.4	2.14	2.15
0.9	0.4	2.16	2.17
0.98	0.4	2.18	2.19

Tabuľka 2.1: Vybrané parametre experimentu a odkaz na grafické znázornenie výsledku

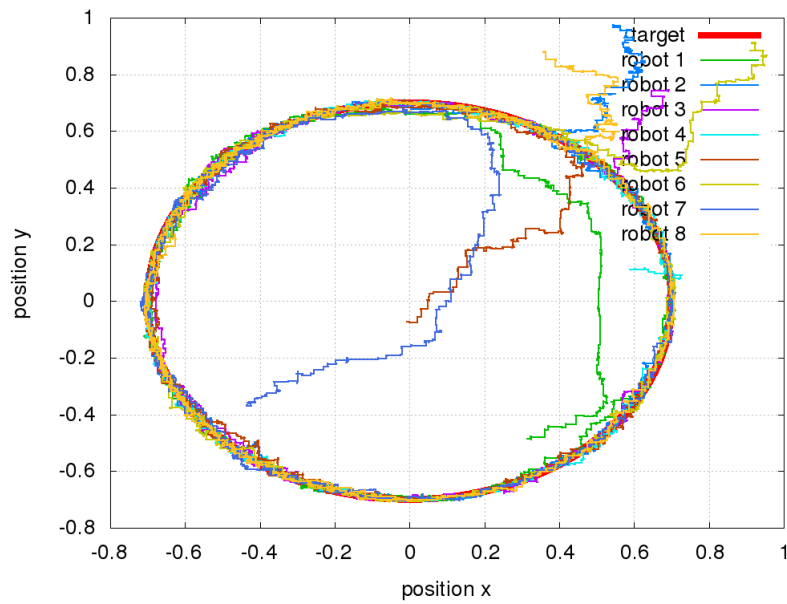


Obr. 2.8: Dráha robotov pre $\gamma = 0.7p = 0.0$

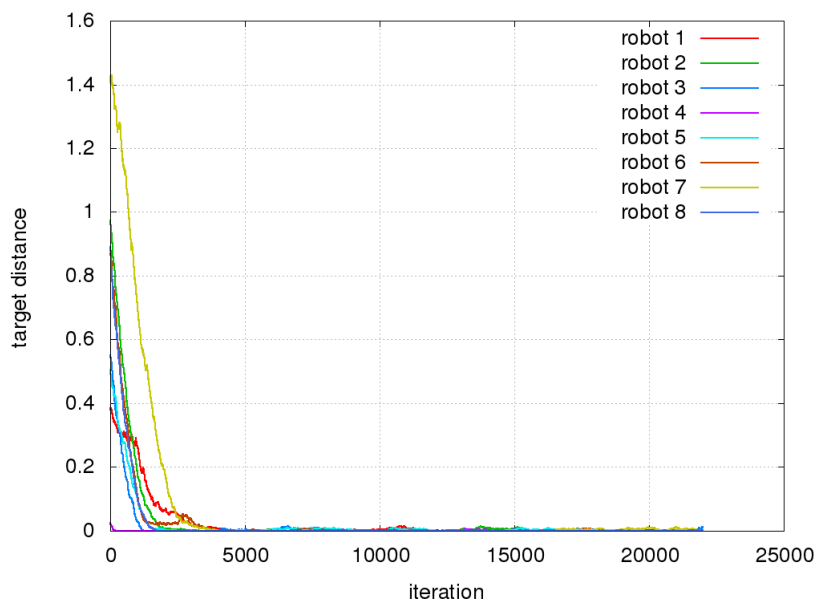


Obr. 2.9: Vzdialenosť robotov od cieľa pre $\gamma = 0.7p = 0.0$

Z experimentov je zrejmé, že pre málo zašumené prostredie nemá hodnota parametra γ veľký význam, obrázky : 2.8 2.9, 2.10 2.11.

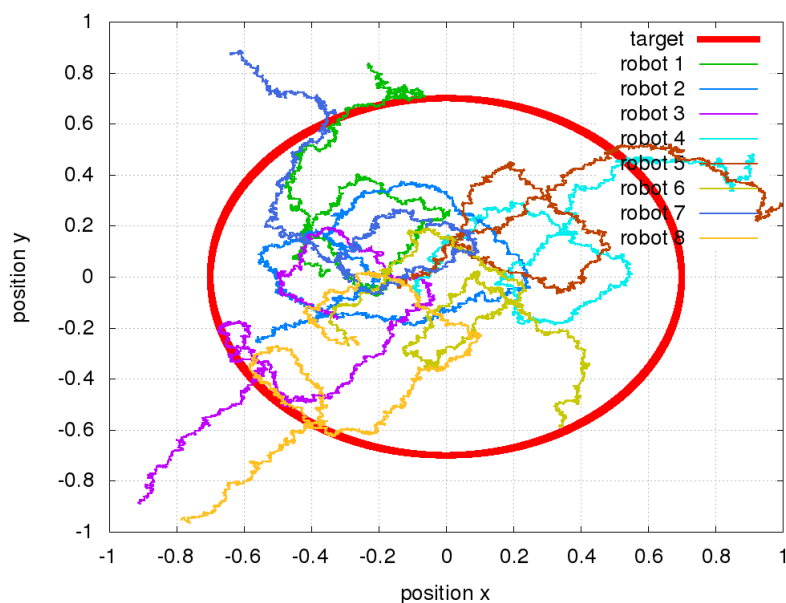


Obr. 2.10: Dráha robotov pre $\gamma = 0.7p = 0.3$

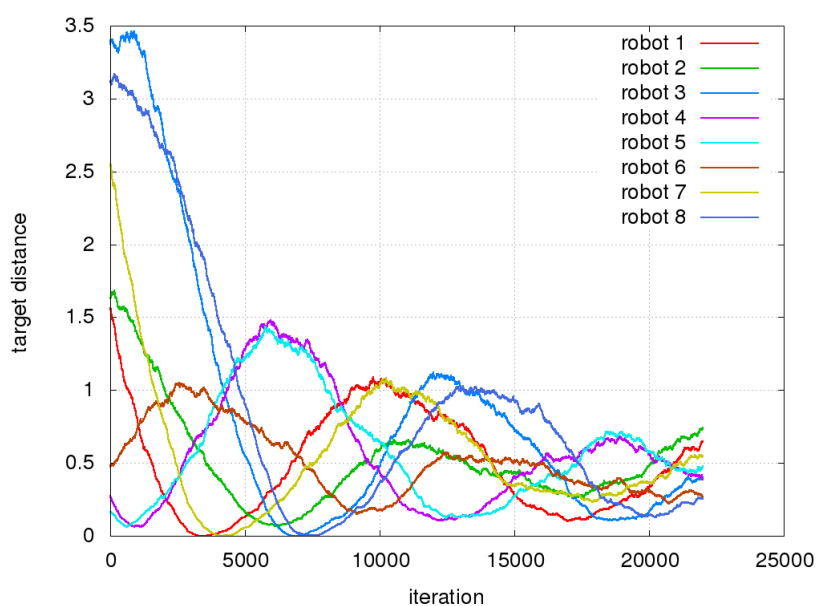


Obr. 2.11: Vzdialenosť robotov od cieľa pre $\gamma = 0.7p = 0.3$

S postupným nárastom šumu však pri nevhodne zvolenej hodnote γ roboti vykazujú veľkú chybu. Je preto tomu potrebné náležite zväčšovať parameter γ , tento priebeh zlepšovania výsledku zvyšovaním parametra γ je zrejmi z obrázkov : 2.12 2.13, 2.14 2.15, 2.16 2.17, 2.18 2.19.

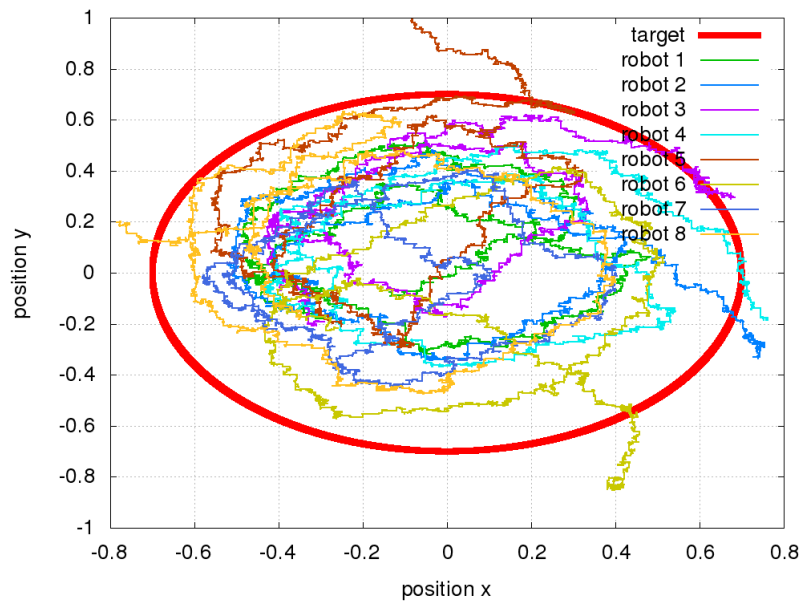


Obr. 2.12: Dráha robotov pre $\gamma = 0.0$ $p = 0.4$

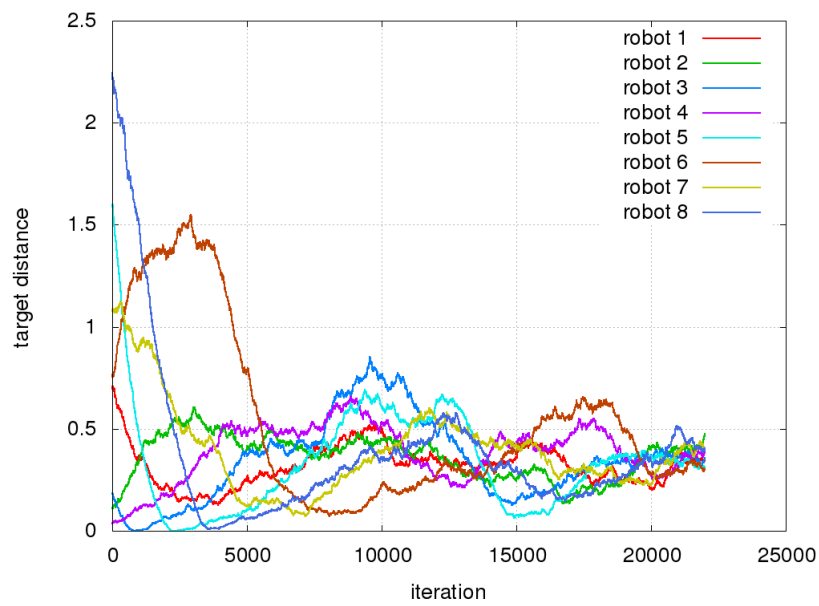


Obr. 2.13: Vzdialenosť robotov od cieľa pre $\gamma = 0.0$ $p = 0.4$

Algoritmus bol pomenovaný nanoQ a je pod licenciou GNU GPL dostupný na [1]. Je napísaný v jazyku C len s využitím pevnej rádovej čiarky. To umožňuje jeho implementáciu aj do menej výkonných mikrokontrolérov, s jadrami napr. Cortex M0 alebo MSP430. Učenie prebieha v reálnom čase a nevyžaduje veľký výpočtový výkon. Podobne, aj pamäťové nároky

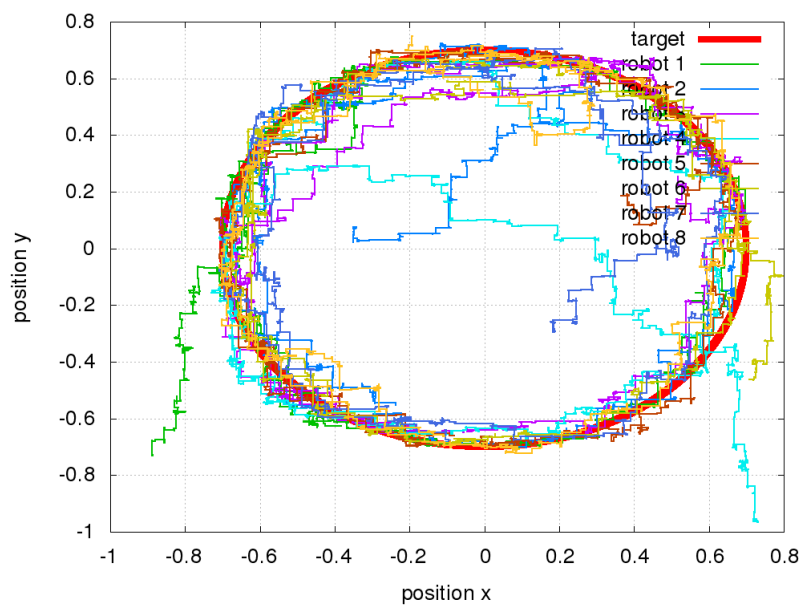


Obr. 2.14: Dráha robotov pre $\gamma = 0.7p = 0.4$

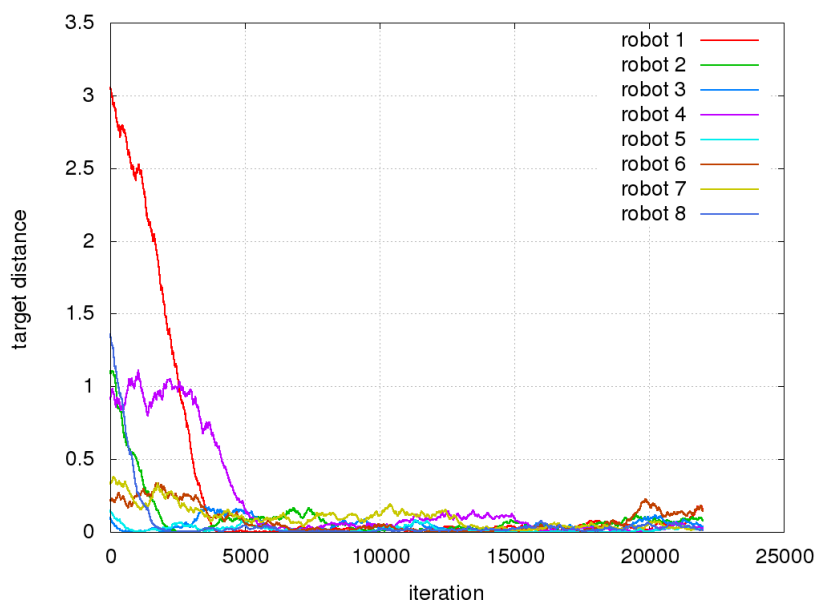


Obr. 2.15: Vzdialenosť robotov od cieľa pre $\gamma = 0.7p = 0.4$

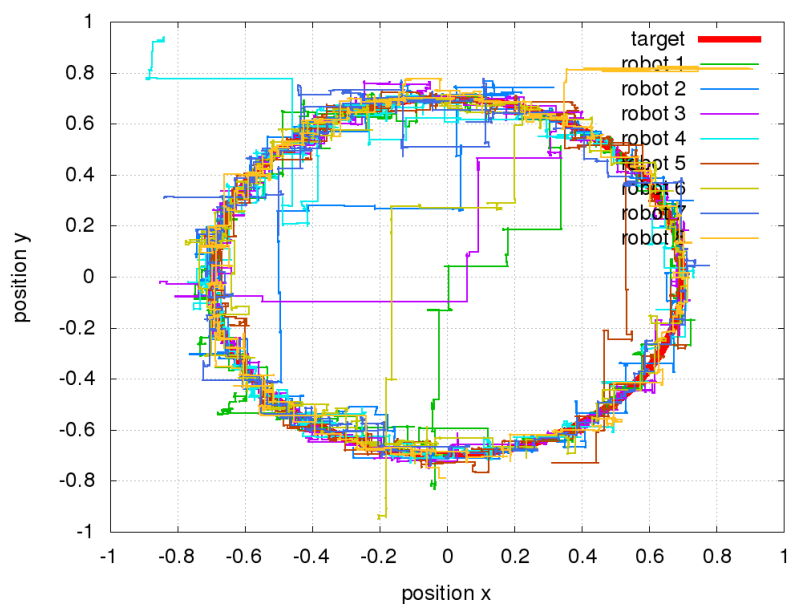
rastú lineárne s počtom akcií. Pre veľký počet akcií je vždy možné použiť aproximáciu, ktorá bude rozobratá v neskorších častiach práce.



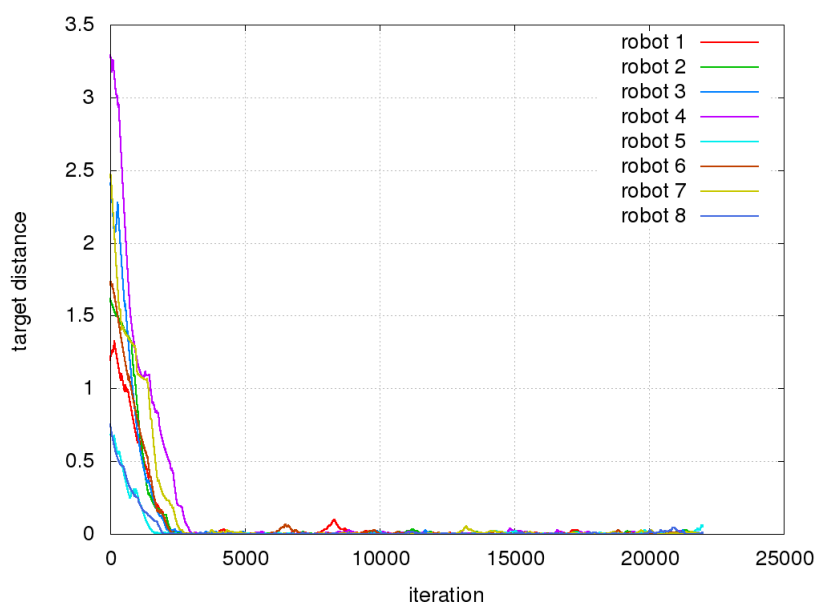
Obr. 2.16: Dráha robotov pre $\gamma = 0.9p = 0.4$



Obr. 2.17: Vzdialenosť robotov od cieľa pre $\gamma = 0.9p = 0.4$



Obr. 2.18: Dráha robotov pre $\gamma = 0.98p = 0.4$



Obr. 2.19: Vzdialenosť robotov od cieľa pre $\gamma = 0.98p = 0.4$

Kapitola 3

Q-learning algoritmus

Q-learning algoritmus je definovaný pre časovo diskkrétne systémy. Agent ktorý prechádza stavový priestor vykonaním niektorej z vopred daných akcií získava za tieto prechody odmeny. Cieľom algoritmu je ohodnotiť všetky akcie v jednotlivých stavoch, tak aby bol dosiahnutý ustálený stav a v každom stave bolo možno vybrať akciu prinášajúcu najväčšiu odmenu, v globálnom zmysle.

3.1 Definícia algoritmu

Daná je množina stavov \mathbb{S} a akcií \mathbb{A} , kde $\mathbb{S} \in \mathbb{R}^{n_s}$ a $\mathbb{A} \in \mathbb{R}^{n_a}$, kde n_s a n_a sú počty prvkov stavového vektora a vektora akcií.

Existuje prechodová funkcia

$$s(n+1) = \lambda(s(n), a(n)) \quad (3.1)$$

zo stavu $s(n) \in \mathbb{S}$ použitím akcie $a(n) \in \mathbb{A}$, táto funkcia je ale algoritmu neznáma.

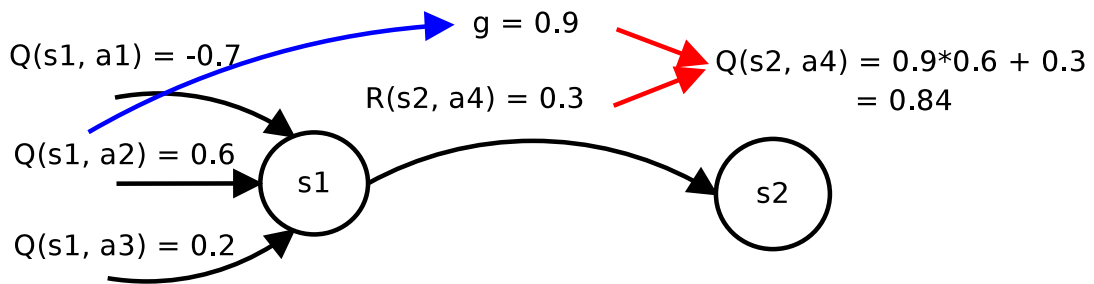
Ďalej je daná odmeňovacia funkcia $R(s(n), a(n))$, ktorá vyjadruje okamžité ohodnotenie konania agenta v $s(n)$ a $a(n)$. V reálnych aplikáciach táto funkcia nadobúda takmer v každom $s(n)$ a $a(n)$ hodnotu 0. Pre správnu funkciu algoritmu, musí byť aspoň jedna hodnota nenulová - napr. ohodnotenie dosiahnutia cieľového stavu (samotná existencia cieľového stavu však pre algoritmus nie je potrebná).

Funkcia ohodnotení je definovaná ako

$$Q_n(s(n), a(n)) = R(s(n), a(n)) + \gamma \max_{a(n-1) \in \mathbb{A}} Q_{n-1}(s(n-1), a(n-1)) \quad (3.2)$$

- $R(s(n), a(n))$ je odmeňovacia funkcia
- $Q_{n-1}(s(n-1), a(n-1))$ je funkcia ohodnotení v stave $s(n-1)$ pre akciu $a(n-1)$
- γ je odmeňovacia konštanta a platí $\gamma \in (0, 1)$.

Funkcia 4.1 definuje ohodnotenie akcií vo všetkých stavoch t.j. agent ktorý sa dostal do stavu $s(n)$ vykonaním akcie $a(n)$ zo stavu $s(n-1)$ získal odmenu $R(s(n), a(n))$ a zlomok najväčšieho možného ohodnotenia ktoré mohol získať dostaním sa do stavu $s(n-1)$, situáciu ilustruje obrázok 3.1.



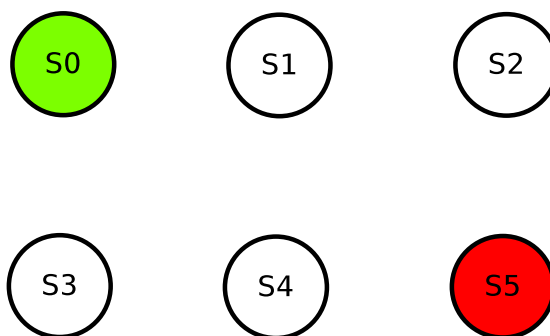
Obr. 3.1: Ilustrácia funkcie ohodnotení, pre $\gamma = 0.9$

Nasledujúce obrázky ilustrujú beh algoritmu pre systém so 6 stavmi pre $\gamma = 0.8$. Na začiatku nie sú známe ani samotné prechody medzi stavmi (Obr. 3.2), bol definovaný 1 cieľový stav $S5$, agent začína v stave $S0$ (môže však v ľubovoľnom inom). Ďalej sa pre jednoduchosť predpokladá že

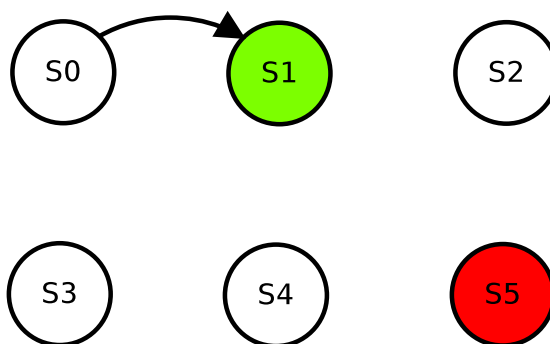
$$R(s(n), a(n)) = \begin{cases} 1 & \text{ak } s(n) = S5 \\ -0.5 & \text{ak } s(n) = S4 \wedge a(n) = A_y \\ 0 & \text{inak} \end{cases} \quad (3.3)$$

t.j. odmeňovacia funkcia nadobúda hodnotu 1 len ak sa agent dostal do stavu $S5$ a pre ilustráciu je definovaná aj jedna záporná odmena pri prechode z $S4$ do $S3$ akciou A_y .

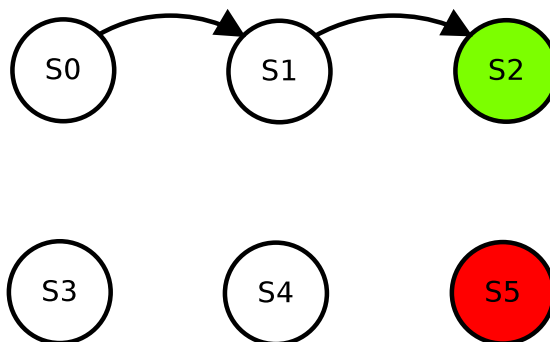
Agent v každom stave náhodne vyberá akcie (na výbere nezáleží, dôležité je aby každá akcia mala nenulovú pravdepodobnosť výberu, a rovnako bola nenulová pravdepodobnosť dosiahnutia ľubovoľného stavu). Obrázky Obr. 3.3 a Obr. 3.4 ilustrujú jednu z možných ciest.



Obr. 3.2: Inicializácia



Obr. 3.3: Prechod do stavu S1

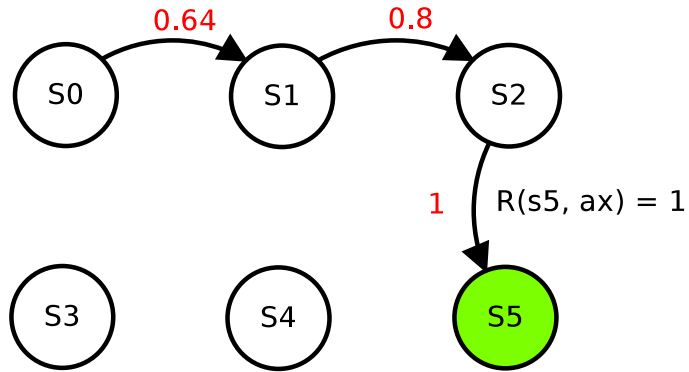


Obr. 3.4: Prechod do stavu S2

Po dosiahnutí cieľového stavu Obr. 3.5 je na základe 3.3 možné spočítať podľa 4.1 ohodnotenia doteraz vykonaných akcií - agent získal nenulovú odmenu $R(s_5, a_x) = 1$ (kde a_x značí ľubovoľnú akciu), ktorú rekurentne spočíta pre všetky doteraz vykonané akcie.

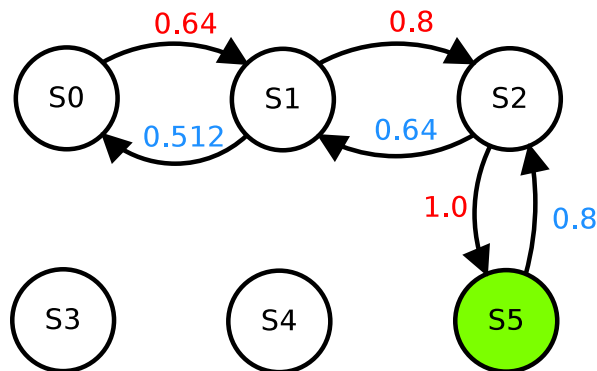
Pre zjednodušenú variantu 4.1 by bolo možné pamätať si len jeden predošlý stav a nepostupovať v ohodnocovaní rekurentne. V praktickej aplikácii je potrebné obmedziť hĺbku rekurzie, a pamätať si len posledných P stavov a v nich urobených rozhodnutiach. V tomto

jednoduchom príklade však nie sú nutné tieto obmedzenia, je teda možné pamätať si celú cestu.



Obr. 3.5: Prechod do stavu S3

Agent môže pokračovať v ceste ďalej, napr. späť 3.6 a približne počítat' ohodnotenia. Prechod z s_5 do s_2 je ohodnotený ako 0.8 - vybralo sa najlepšie možné ohodnotenie ako sa dostať do s_5 (1) násobené γ , $R(s_2, a_x) = 0$ (podľa 3.3).

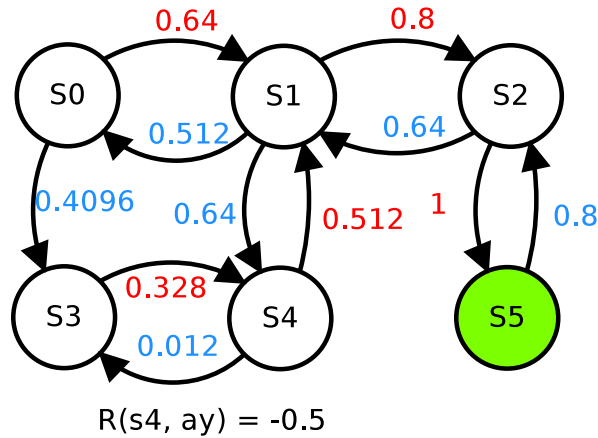


Obr. 3.6: Ďalšie prechody agenta

Po prejdení celého grafu, kedy agent vykonal všetky možné akcie dosiahne funkcia $Q(s(n), a(n))$ konečný, ustálený stav Obr. 3.7, teda

$$\forall s(n), \forall a(n), \forall \varepsilon > 0 \exists Q_n : |Q_n(s(n), a(n)) - Q_{n-1}(s(n), a(n))| < \varepsilon \quad (3.4)$$

hodnoty Q-funkcie sa teda pri pevne danom $R(s(n), a(n))$ už nemenia.



Obr. 3.7: Konečný stav

3.2 Výber akcie

Pre nájdenie konečných hodnôt funkcie ohodnotení podľa 3.4 stačí aby každý prechod mal nenulovú pravdepodobnosť vykonania. Pre ďalšie vyšetřovanie konania agenta je daná pravdepodobnosť výberu akcie ako

$$P(s(n), a(s)) = \frac{e^{kQ(s(n), a(s))}}{\sum_{i=1}^{C_a} e^{kQ(s(n), a(i))}} \quad (3.5)$$

kde

s je zvolená akcia

C_a je počet akcií

k je konštanta a platí $k \geq 0$

agent ktorý vyberá všetky akcie s rovnakou pravdepodobnosťou má teda $k = 0$. Pre vysoké hodnoty $k : \lim_{k \rightarrow \infty}$ bude agent vyberať len najlepšie dostupné akcie. Pre učenie agenta je teda vhodné zvoliť malé k .

Je možné definovať ľubovoľné iné možnosti výberu akcie, napr. uprednostňovať menej často vykonané akcie, prípadne podľa zmeny $|Q_n(s(n), a(n)) - Q_{n-1}(s(n), a(n))|$ uprednostňovať prechody s veľkou hodnotou zmeny.

3.3 Problémy výpočtu $Q(s, a)$

Algoritmus je definovaný pre diskretnú množinu stavov. Ďalej sa predpokladá, že $s(n) \in \langle -1, 1 \rangle$ a podobne $a(n) \in \langle -1, 1 \rangle$

Pre počty prvkov stavového vektora a vektora akcií (n_s, n_a) je možné definovať delenie ich hodnôt na diskretný počet d_s a d_a , potom je možné vyjadriť celkový počet hodnôt $Q(s(n), a(n))$ ako

$$C = d_s^{n_s} d_a^{n_a} \quad (3.6)$$

Samotný počet hodnôt ktoré treba spočítať teda exponenciálne narastá s rastom počtu prvkov stavového a vektora akcií.

Pre úlohy kde do systému vstupuje mnoho nezávislých vstupov sa stáva implementácia $Q(s(n), a(n))$ problémom najmä z dôvodov :

- veľké pamäťové nároky
- o nenavštívených prechodoch nevie agent povedať nič

Vhodným riešením sa ukazuje aproximácia Q-funkcie. Nech je aproximovaná funkcia označená $Q'_n(s(n), a(n))$ a presné riešenie ako $Q_n(s(n), a(n))$.

Dané sú postuláty o tejto aproximácii

Postulát 1 *Neobmedzená prenosť aproximácie* : Pre všetky stavy $s(n)$ a akcie $a(n)$ musí platiť $|Q_n(s(n), a(n)) - Q'_n(s(n), a(n))| < \varepsilon$. Kde $\varepsilon > 0$ a určuje kvalitu aproximácie. Zlepšením vlastností $Q'_n(s(n), a(n))$ je možné ľubovoľne znižovať ε .

Postulát 2 *Lokálna zmena* : Lokálna zmena hodnoty $\delta = |Q'_n(s(n), a(n)) - Q'_{n-1}(s(n), a(n))|$ neovplyvní hodnotu funkcie v inom bode o viac ako $\forall s(n') \forall a(n'), n \neq n' : \delta < \kappa$. Znižovaním hodnoty κ sa funkcia stáva menej závislá na okolí bodu $[s(n), a(n)]$.

Funkciu $Q(s(n), a(n))$ je možné aproximovať niekoľkými spôsobmi. Tie najbežnejšie sú

- tabuľka

- neurónová sieť
 - dopredná neurónová sieť
 - kohonenova mapa
 - neurónová sieť bázičických funkcií

3.4 Tabuľka

Priamočiarým prístupom je ukladať hodnoty $Q(s(n), a(n))$ do tabuľky. Pre diskretnú a konečnú množinu akcií možno tabuľku rozdeliť na viac častí, čím sa urýchli vyhľadávanie. Počet stavov v reálnych aplikáciach však môže byť vysoký (vo všeobecnosti tvoria stavy spojitú množinu).

Stav agenta aj vykonaná akcia sú obvykle vektory normované do $\langle -1, 1 \rangle$, pre implementáciu tabuľky je nutné ich prepočítať na celočíselné indexy

$$I_s(n) = \lceil \sum_{i=1}^{n_s} \left(d_s \frac{s_i(n) + 1}{2} \right)^i \rceil \quad (3.7)$$

$$I_a(n) = \lceil \sum_{i=1}^{n_a} \left(d_a \frac{a_i(n) + 1}{2} \right)^i \rceil \quad (3.8)$$

kde

$I_s(n)$ je index stavu

$I_a(n)$ je index akcie

$s_i(n)$ je i -ty prvok vektora stavu $s(n)$

$a_i(n)$ je i -ty prvok vektora akcií $a(n)$

Pre diskretný počet stavov a akcií je možné definovať tabuľkovú interpretáciu ako $Q^t(I_s(n), I_a(n))$. Pre $\lim_{d_s \rightarrow \infty}$ a $\lim_{d_a \rightarrow \infty}$ je možné považovať tabuľku za presné riešenie pretože spĺňa postuláty 1 aj 2.

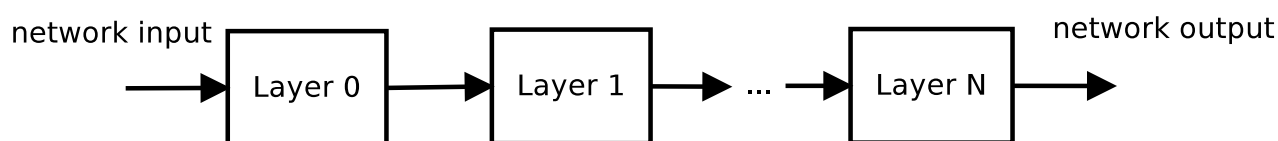
3.5 Dopredná neurónová sieť

Pre aproximáciu funkcie ohodnotení je možné použiť dobrednú neurónovú sieť ako univerzálny aproximátor. Podľa Kolmogorovho teorému je možné neurónovú sieť takto použiť [2], [3] a [4]. Samotný teorém však nerieši problém učenia takejto siete. Na učenie siete existuje niekoľko algoritmov, medzi najčastejšie patria [5], [6], [7] :

1. Backpropagation
2. Genetické algoritmy
3. Simulované žihanie

Najpoužívanejší Backpropagation má množstvo problémov, najmä postupný pokles gradientu v hlbokých neurónových sieťach. Kedy sú vnorené vrstvy prakticky neučené. Ďalší závažný problém je uviaznutie v lokálnom minime, čo sa autory snažia vyriešiť zavedným zotrvačnosťou do zmeny váh siete. Istým nádejším východiskom sa zdá byť simulované žihanie [8]. Pre problém Q-learning algoritmu je však potrebný prístup ktorý postupne zlepšuje riešenie, a to je jedine niektorá z gradientových metód.

Samotná dopredná sieť je tvorená niekoľkými prepojenými vrstvami neurónov 3.8. Pre popis vrstvy je však vhodné použiť maticový zápis, nakoľko prenos neurónu má obvykle vo vrstve rovnakú aktivačnú funkciu.



Obr. 3.8: Dopredná neuronová sieť

Správanie sa jednej vrstvy je možné popísať ako funkciu vstupného vektora, ktorej výstupom je tiež vektor (počty prvkov týchto vektorov však môžu byť rôzne, vždy však konečné). Je daný vstupný vektor

$$I(n) = (s(n), a(n)) \quad (3.9)$$

výstupná hodnota neurónovej siete ako $y_{nn}(n)$ a požadovaná hodnota ako $y_r(n)$.

Ďalej je definovaná chyba ako

$$e(n) = y_r(n) - y_{nn}(n) \quad (3.10)$$

Vrstva l doprednej siete je definovaná ako

$$\begin{aligned} y^l(n) &= f^l \left(W^l(n) I^l(n) \right) \\ &= f^l \left(\begin{pmatrix} w_{1,1}^l(n) & w_{1,2}^l(n) & \cdots & w_{1,n'}^l(n) \\ w_{2,1}^l(n) & w_{2,2}^l(n) & \cdots & w_{2,n'}^l(n) \\ \vdots & \vdots & \ddots & \vdots \\ w_{m',1}^l(n) & w_{m',2}^l(n) & \cdots & w_{m',n'}^l(n) \end{pmatrix} \begin{pmatrix} i_{1,1}^l(n) & i_{1,2}^l(n) & \cdots & i_{1,n'}^l(n) \end{pmatrix} \right) \end{aligned} \quad (3.11)$$

kde

n' je počet prvkov vstupného vektora

m' je počet prvkov výstupného vektora

$f(X)$ je aktivačná funkcia

$W^l(n)$ je matica váh.

Najčastejšie používané aktivačné funkcie sú sigmoida, hyperbolický tangens, lineárna, usmerňovač a skoková funkcia. Ich predpisy sú

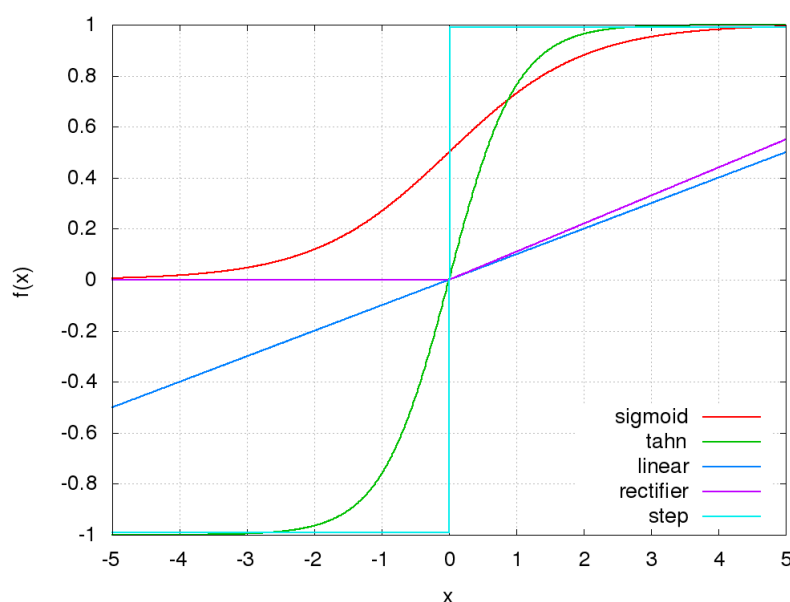
$$y_1(x) = \frac{1}{1 + e^{-x}} \quad (3.12)$$

$$y_2(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3.13)$$

$$y_3(x) = x \quad (3.14)$$

$$y_3(x) = \begin{cases} x & \text{ak } x > 0 \\ 0 & \text{inak} \end{cases} \quad (3.15)$$

$$y_4(x) = \begin{cases} 1 & \text{ak } x > 0 \\ -1 & \text{inak} \end{cases} \quad (3.16)$$



Obr. 3.9: Grafické znázornenie priebehov aktivačných funkcií

Ich priebehy sú znázornené na obrázku 3.9.

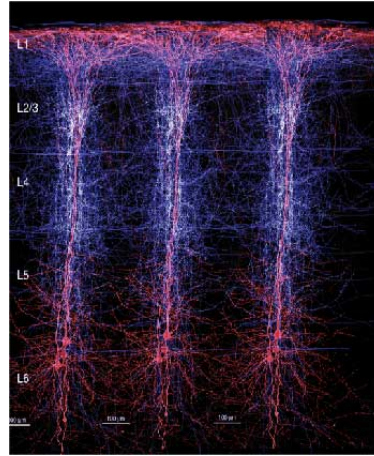
Zoradením niekoľkých vrstiev za sebou, tak že výstup predošlej je vstupom do aktuálnej vrstvy je možné získať doprednú neurónovú sieť. Takáto sieť je vhodná na riešenie klasifikačných aj aproximačných problémov.

Dopredná sieť je známa nelokálnosťou učenia : pri tréňovaní na podmnožinu množiny požadovaných výstupov sa mení hodnota aj mimo túto podmnožinu. Sieť teda veľmi problematcky spĺňa postulát 2. Dobré však spĺňa postulát 1, vhodnou voľbou počtu vrstiev a počtu neurónov je možné dosiahnuť ľubovoľnú presnosť aproximácie.

3.6 Kohonenová neurónová sieť

Kohonenova neurónová sieť je inšpirovaná rovinnými štruktúrami v mozgovej kôre, kde výrazne dominujú prepojenia v rámci vrstvy a prepojenia medzi vrstvami je podstatne menej 3.10. Napriek týmto rovinatým štruktúram, mozog bežne spracúva mnohodomenzionálne problémy - vstupom sú podnety s miliónov receptorov.

Pôvodný návrh Kohonenovej siete neriešil aproximáciu funkcie, ale klasifikáciu vstupov do zhukov. Pričom podobné vstupy patria do rovnakého zhuku. Počet zhukov je voliteľný



Obr. 3.10: Neokortexový stĺpec

a je rovný počtu použitých neurónov. Takáto sieť sa vie učiť bez definovania chyby, t.j. nie je potrebné stanoviť do ktorého zhluku dáta patria. Samotné zhluky sa utvárajú postupne, tak ako sú do siete predkladané dáta.

Každému zhluku je však možné priradiť požadovanú výstupnú hodnotu. Sieť tak môže pracovať podobne ako tabuľka - vstupné dáta priradí do najbližšieho zhluku a výstupom je hodnota prisluchujúca k tomuto zhluku. V tomto prípade už treba stanoviť chybu, ako rozdiel požadovanej hodnoty a skutočnej hodnoty výstupu siete. Algoritmus má niekoľko krokov : Najpr sa spočítajú vzdialenosti od vstupného vektora

$$d_j(n) = \sum_{i=1}^N (I_i(n) - w_{ji}(n))^2 \quad (3.17)$$

kde $w(n) \in \mathbb{R}$ je matica váh, na začiatku sa volí náhodná, tak aby rovnomerne pokrila stavový priestor vstupných veličín (Obvykle sa používa rovnomerné rozdelenie. Štatistickou analýzou vstupných dát je ale možné určiť iné, vhodnejšie a urýchliť tak konvergenciu hodnôt váh).

Vít'azný neurón v je definovaný ako

$$v : \forall j : d_v(n) \leq d_j(n) \quad (3.18)$$

je to neurón ktorý má najmenšiu vzdialenosť od predloženého vstupu.

A pre každý neurón existuje priradená výstupná hodnota $y_j(n) \in \mathbb{R}$, výstupom siete je teda hodnota priradená víťaznému neurónu $y_{nn}(n) = y_v(n)$.

Učenie siete prebieha v dvoch krokoch

1) **zmena váh** $w(n)$ zmenia sa váhy víťazného neurónu, pretože najlepšie zodpovedajú požadovaným váham tak že sa priblížia hodnote vstupného vektora

$$w_{ji}(n+1) = (1 - \eta_1(j))w_{ji}(n) + \eta_1 I_i(n) \quad (3.19)$$

kde $\eta_1(j) \in (0, 1)$ je krok učenia a závisí od polohy neurónu v sieti. V najjednoduchšom prípade

$$\eta_1(j) = \begin{cases} \eta & \text{ak } j = v \\ 0 & \text{inak} \end{cases} \quad (3.20)$$

k zmene váh teda dôjde len pri víťaznom neuróne. Ďalší často používaný tvar funkcie postupne znižuje hodnotu $\eta_1(j)$ podľa $d_j(n)$ a to ako

$$\eta_1(j, n) = \eta e^{-kd_j(n)} \quad (3.21)$$

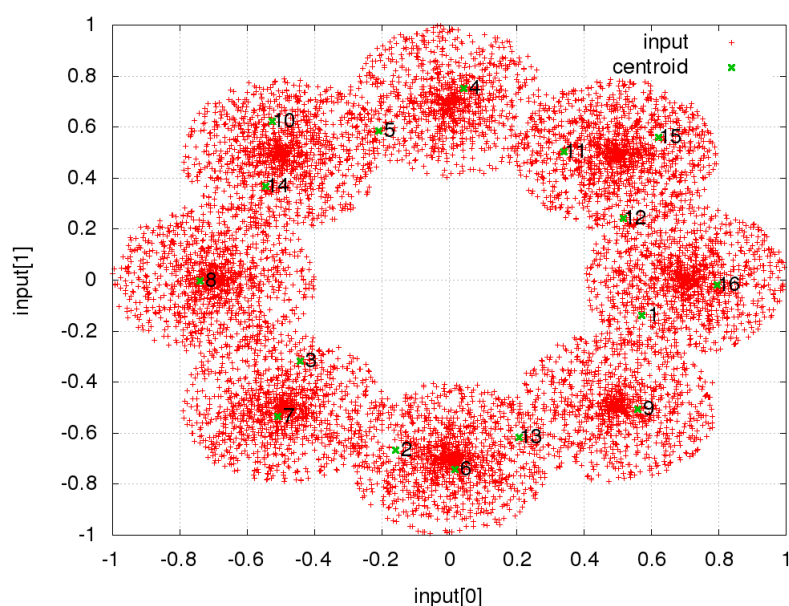
kde $k \in (0, \infty)$. Krok učenia $\eta_1(j, n)$ je teda premenný a závisí aj od predloženého vzoru podľa n .

Po dostatočnom počte iterácií sa hodnoty váh ustália na hodnotách tak aby rozdelili množinu vstupných vektorov na lokálne oblasti. Tento stav je znázornený na 3.11. Vstupný proces generoval 8 zhlukov dát ktoré sieť klasifikovala použitím 16 neurónov. Každému neurónu je možné priradiť požadovanú hodnotu výstupu.

2) **upraví sa výstupná hodnota** $y_v(n)$

$$y_v(n+1) = y_v(n) + \eta_2 e(n) \quad (3.22)$$

kde $e(n)$ je chyba podľa 3.10. Takáto Kohonenová sieť má veľmi dobré predpoklady na aproximáciu funkcie, ktorá nadobúda málo hodnôt, predkladaním vzorov z blízkosti jedného



Obr. 3.11: Znázornenie váhových parametrov w pre dvojrozmerný priestor

zhľuku (dáta majú malú vzdialenosť v zmysle 3.17) sa dá očakávať zmena parametrov len jedného neurónu. Sieť teda dobre spĺňa predpoklady 1, 2. Počet neurónov však musí narásť do nekonečna.

Je vhodné poznamenať, že výstupná hodnota nemusí byť len číselná hodnota, ale aj funkcia, ktorej vstupom je opäť $I_i(n)$. Výber výťazného neurónu tak funguje ako výber vhodnej funkcie - prepínač.

3.7 Neurónová sieť bázických funkcií

Samotný prenos neurónu nemusí byť obmedzený len na množinu funkcií z 3.16. Vhodnú funkciu je možné zmenou parametrov upraviť do tvaru, aby na zvolený vstup $I_0(n)$ dosahovala požadovanú hodnotu a postupným zväčšovaním vzdialenosti $|I_0(n) - I_i(n)|$ klesala jej hodnota k nule.

Najjednoduchším príkladom takýchto funkcií je

$$f_j(X(n)) = \begin{cases} k_j & \text{ak } X(n) = X_0 \\ 0 & \text{inak} \end{cases} \quad (3.23)$$

kde k_j je hodnota požadovaná v bode X_0 . Výstupom siete potom je

$$y(X) = \sum_{j=1} f_j(X(n)) \quad (3.24)$$

Z charakteru Q-learning algoritmu majú hodnoty $Q(s(n), a(n))$ charakter aj postupne klesajúcich hodnôt. Je teda potrebné vybrať iné funkcie.

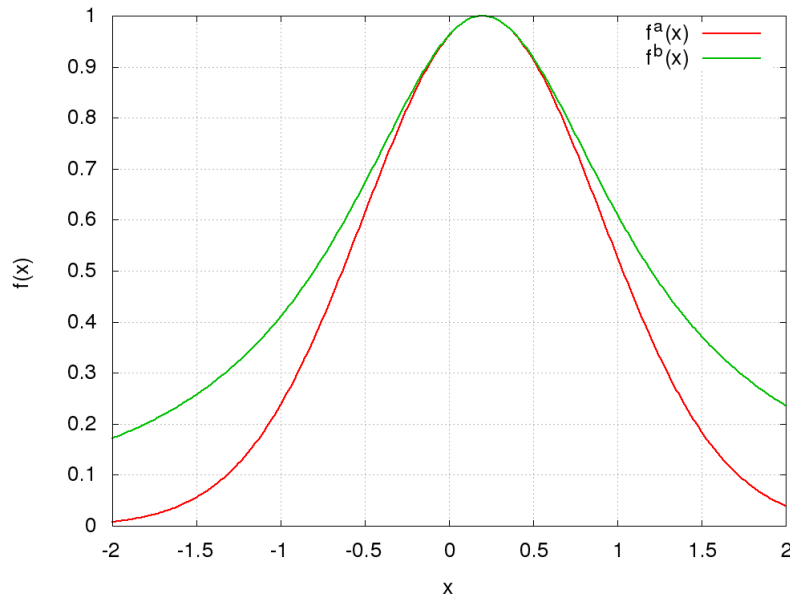
Nasledujú preto definície funkcií s ktorými boli urobené experimenty.

Dané sú základné funkcie $f_j^x(s(n), a(n))$, kde x je typ základnej funkcie. Požadovaná hodnota $Q^x(s(n), a(n))$ je potom lineárnou kombináciou týchto funkcií typu x .

Z charakteru Q-learning algoritmu 4.1 je možné určiť požiadavky na tieto funkcie :

1. predpis 4.1 je tvorený klesajúcou exponenciálou - podobný charakter by mala mať aj základná funkcia
2. existencia jedného globálneho maxima a zmenou parametrov určovať polohu tohto bodu
3. možnosť ľubovoľne meniť strmosť funkcie v okolí maxima
4. funkcia by mala byť zhora aj z dola ohraničená

Cieľom je mať možnosť nezávisle nastaviť maximá funkcií do oblastí, ktoré zodpovedajú nenulovým hodnotám $R(s(n), a(n))$ - bod 2. Ak ohodnotenie spĺňa podmienku najlepšej možnej akcie v danom stave, dá sa očakávať že bude mať menšiu strmosť, naopak, ak funkcia popisuje bod kde $R(s(n), a(n))$ dosahuje malé hodnoty (obvykle záporné), bude požadovaná vysoká strmosť tejto funkcie - obe požiadavky sú zhrnuté v bode 3. Bod 4 umožňuje rozumne ohraničiť rozsah funkcie.



Obr. 3.12: Znázornenie priebehov bázických funkcií

Niektoré tvary bázických funkcií

$$f_j^1(s(n), a(n)) = e^{-\sum_{i=1}^{n_s} \beta_{aji}(n)(s_i(n) - \alpha_{aji}(n))^2} \quad (3.25)$$

$$f_j^2(s(n), a(n)) = \frac{1}{1 + \sum_{i=1}^{n_s} \beta_{aji}(n)(s_i(n) - \alpha_{aji}(n))^2} \quad (3.26)$$

$$f_j^3(s(n), a(n)) = e^{-\sum_{i=1}^{n_s} \beta_{aji}(n)|s_i(n) - \alpha_{aji}(n)|} \quad (3.27)$$

kde

$\alpha_{aji}(n) \in \langle -1, 1 \rangle$ určuje polohu maxima funkcie

$\beta_{aji}(n) \in (0, \infty)$ určuje strmosť funkcie.

Ich priebeh pre prvé dve je na 3.12.

Pre symetrické prechody medzi stavmi ich možno zjednodušiť na

$$f_j^1(s(n), a(n)) = e^{-\beta_{aj} \sum_{i=1}^{n_s} (s_i(n) - \alpha_{aji})^2} \quad (3.28)$$

$$f_j^2(s(n), a(n)) = \frac{1}{1 + \beta_{aj} \sum_{i=1}^{n_s} (s_i(n) - \alpha_{aji})^2} \quad (3.29)$$

$$f_j^3(s(n), a(n)) = e^{-\beta_{aj} \sum_{i=1}^{n_s} |s_i(n) - \alpha_{aji}|} \quad (3.30)$$

Aproximovaná funkcia ohodnotení pre l bázičských funkcií je potom

$$Q^x(s(n), a(n)) = \sum_{j=1}^l w(n)_j^x f_j^x(s(n), a(n)) \quad (3.31)$$

kde $w(n)_j^x$ sú váhy bázičských funkcií.

Je teda potrebné stanoviť celkovo 3 sady parametrov : α β w .

3.7.1 Určenie parametrov α

Parameter α určuje posunutie maxima funkcie a postupuje sa podobne ako v prípade 3.19. Treba zohľadniť fakt, že pre konečný výsledok je dôležité pokryť všetky oblasti s nenulovým $R(s(n), a(n))$, vrchol krivky bude ležať nad bodom $[s(n), a(n)]$.

Zmena parametrov α prebieha v piatich krokoch.

- na začiatku sa zvolia $\alpha_{jia}(n)$ náhodne, ze $\langle -1, 1 \rangle$
- spočítajú sa vzdialenosti od predloženého vstupu $d_{ja}(n) = |s(n) - \alpha_{ja}(n)|$
- nájde sa také ka kde pre $\forall j : d_{ka}(n) \leq d_{ja}(n)$
- spočíta sa krok učenia $\eta'_a(n) = \eta_1 |Q_r(s(n), a(n))|$
- upraví sa parametre $\alpha_{aki}(n+1) = (1 - \eta')\alpha_{aki}(n) + \eta' s_i(n)$

kde

$Q_r(s(n), a(n))$ je požadovaný výstup

η_1 je konštanta učenia

Krok učenia teda závisí od veľkosti požadovanej hodnoty, tým sa zabezpečí aby maximum krivky naozaj ležalo nad bodom $[s(n), a(n)]$.

3.7.2 Určenie parametrov β

Parameter β určuje strmosť krivky. Ak boli k dispozícii naraz všetky požadované výstupy, bolo by možné spočítať tento parameter z rozptylu. Požadované hodnoty však prichádzajú postupne, strmosť krivky sa preto upravuje priebežne, podľa toho či požadovaná hodnota leží nad, alebo pod krivkou.

- stanoví sa chyba $e(n) = Q_r(s(n), a(n)) - Q(s(n), a(n))$
- pre každú bázickú funkciu $\beta_{ja}(n+1) = \beta_{ja}(n) + \eta_2 e(n) w_{ja}(n)$
- skontroluje sa $\beta_{ja}(n) \in (0, \infty)$

kde

$Q_r(s(n), a(n))$ je požadovaný výstup

η_2 je konštanta učenia

3.7.3 Určenie váhových parametrov w

Nakoniec sa gradientovou metódou určia váhové parametre. Pre presné riešenie by bolo možné použiť metódu najmenších štvorcov, tá je však pre veľký počet bázických funkcií ťažko vypočítateľná. Zmena parametrov je potom daná nasledujúcim postupom

- stanoví sa chyba $e(n) = Q_r(s(n), a(n)) - Q(s(n), a(n))$
- pre každé $w_{ja} : w_{ja}(n+1) = w_{ja}(n) + \eta_3 e(n) y_j(n)$
- skontroluje sa $w_{ja}(n) \in (-r, r)$

kde

η_3 je konštanta učenia

r je maximálny rozsah váh

3.7.4 Hybridný variant

Ak by funkcia $R(s(n), a(n))$ mala len jednu kladnú hodnotu a ostatné by boli nulové, aproximáciu $Q(s(n), a(n))$ by veľmi dobre popísala Gaussova krivka 3.30. Ak by funkcia $R(s(n), a(n))$ mala len záporné hodnoty a ostatné by boli nulové, funkcia $Q(s(n), a(n))$ by s ohľadnutím na 4.1 by si boli rovné. Vo funkcií $Q(s(n), a(n))$ by sa tak objavilo niekoľko záporných hodnôt, ostro ohraničených.

Vyjdúc z týchto úvah, je možné skombinovať výhody oboch : Gaussova krivka ktorá dokáže pokryť nenulovými hodnotami celý definíčný obor a funkcie 3.23.

Je teda možné skombinovať funkciu 3.23 s niektorou z 3.30, čo vedie na vzťahy

$$H_j(s(n)) = \begin{cases} r_j & \text{if } s(n) = \alpha_j \\ 0 & \text{inak} \end{cases} \quad (3.32)$$

$$f_j(s(n), a(n)) = H_j(s(n)) + w_j e^{-\beta_{aj} \sum_{i=1}^{n_s} (s_i(n) - \alpha_{aji})^2} \quad (3.33)$$

$$Q(s(n), a(n)) = \sum_{j=1}^J f_j(s(n), a(n)) \quad (3.34)$$

kde

α_j je stav pre ktorý sa počíta funkcia

r_j je hodnota okamžitej odmeny $R(s(n))$ v tomto stave

β_j je strmosť, a platí $\beta > 0$

w_j je váha

Kapitola 4

Experimentálna časť

4.1 Ciele experimentu

V oblasti Q-learning algoritmov je možné pozorovať dva hlavné smery výskumu

- aproximácia funkcie ohodnotení
- spôsob výberu akcie

Obe majú široké pole diskusií v snahe vyriešiť niekoľko hlavných problém Q-learning algoritmu a to najmä

- veľký počet prechodov medzi stavmi
- malá zmena vo výpočte $Q(s(n), a(n))$ môže spôsobiť veľké zmeny v stratégií.

Cieľom práce je na danej množine odmeňovacích funkcií $R(s(n), a(n))$ overiť možnosti aproximácie $Q(s(n), a(n))$. V niekoľkých bodoch je možné postup určiť ako

- výber funkcií $R(s(n), a(n))$
- určenie presného riešenia, použitím tabuľky s veľkým počtom prvkov
- voľba aproximačnej metódy
- pre každú $R(s(n), a(n))$ spočítať niekoľko nezávislých behov

- výsledky porovnať s presným riešením, overiť a zosumarizovať

Funkcie $R(s(n), a(n))$ budu vybrané tak aby boli riedke a plne sa využil Q-learning - okamžité odmeny sú známe len v malom počte prípadov. Postupne sa obmenia pre rôzne počty nenulových prvkov.

Presné riešenie, aby bolo možné spočítať bude mať niekoľko tisíc diskretných stavov. Pre jednoduchosť, bude v každom stave rovnaká a presne definovaná množina akcií.

Vyberie sa niekoľko aproximačných metód, ktoré sa použijú na spočítanie $Q(s(n), a(n))$. Tu je nevyhnutné upozorniť na častú metodickú chybu : aj keď je možné $Q(s(n), a(n))$ spočítať presne, nesmie byť toto presné riešenie použité na stanovenie približného riešenia. Príkladom je dopredná neurónová sieť, ktorá sa dá veľmi ľahko natréňovať ak je množina požadovaných výstupov vopred známa. V prípade Q-learning algoritmu sa ale požadované hodnoty spočítavajú rekuretné, až počas behu.

Kedže voľba niektorých počiatočných parametrov aproximačných metód je náhodná, je nevyhnutné spočítať niekoľko nezávislých behov a overiť tak rozptyl, minimálnu, maximálnu a priemernu chybu.

4.2 Návrh experimentu

Aby sa dalo kvalitatívne ohodnotiť použité riešenie, je nutné urobiť veľký počet experimentov. Aby bolo možné ľahko graficky znázorniť výsledok, bude stavový priestor dvojrozmerný a platí $s(n) \in \langle -1, 1 \rangle$. Agent si bude vyberať z pevne danej množiny akcií a bude sa tak v tomto priestore môcť pohybovať a to :

$$\mathbb{A} = [[0, 1], [0, -1], [1, 0], [-1, 0], [1, -1], [1, 1], [-1, -1], [-1, 1]]$$

prostredie umožní zmenu stavu vykonaním akcie $a(n) \in \mathbb{A}$, a to podľa

$$s(n+1) = s(n) + a(n)dt \quad (4.1)$$

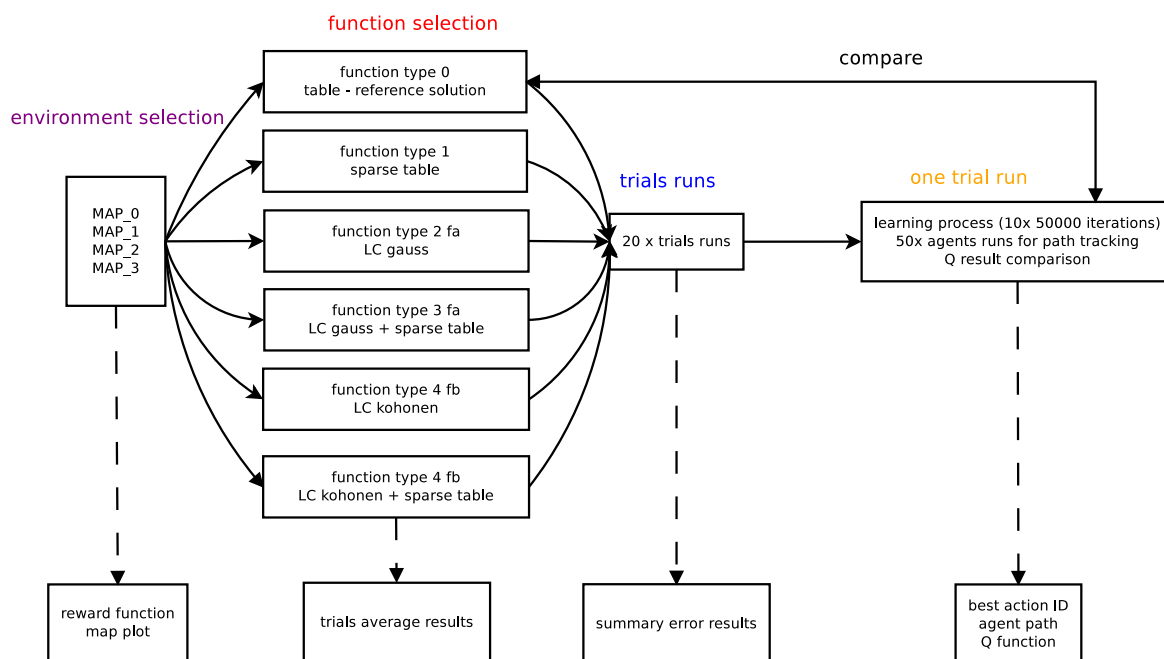
Jednotlivé funkcie $R^k(s(n), a(n))$ predstavujú mapy odmien v ktorých sa agent pohybuje. Pre zjednodušenie bude platiť, že nezáleží ktorou akciou sa agent dostal do daného stavu -

funkcia bude mať teda tvar $R^k(s(n))$ a predstavuje teda odmenu za to, že sa agent dostal na nejaké miesto.

Ako metódy aprximácie je zvolených 5 rôznych funkcií.

1. riedka tabuľka
2. Gaussova krivka $f_j^1(s(n), a(n))$ 3.30
3. Gaussova krivka $f_j^1(s(n), a(n))$ kombinovaná s riedkou tabuľkou
4. Modifikácia Kohonenovej neurónovej siete $f_j^2(s(n), a(n))$
5. Modifikácia Kohonenovej neurónovej siete $f_j^2(s(n), a(n))$ s riedkou tabuľkou

Pre každú z nich prebehne 20 trialov aby bolo možné urobiť štatistické vyhodnotenie. V každom trialu prebehne 10×50000 učiacich interácií aby bolo možné v 10 tich krokoch sledovať priebeh učenia. Na konci prebehne 50 behov agentov z náhodných východných stavov aby bolo možné sledovať ich cestu stavovým priestorom.



Obr. 4.1: Schéma experimentu

Súhrnná schéma behu experimentov je na obrázku 4.1. Plné šípky predstavujú prepojenie úrovni metodológie. Čiarkované šípky znázorňujú výstupy v jednotlivých úrovniach. Presné riešenie je použité na porovnanie výslednej chyby.

- 50000 iterácií učenia
- rozmer s je $n_s = 2$, rozmer a je $n_a = 2$
- predpis funkcie ohodnotení

$$Q(s(n), a(n)) = \alpha Q(s(n-1), a(n-1)) + (1 - \alpha)(R(s(n), a(n)) + \gamma \max_{a(n-1) \in \mathbb{A}} Q(s(n-1), a(n-1)))$$

- $R(s(n), a(n)) \in \langle -1, 1 \rangle$ náhodná mapa s 1 cieľovým stavom
- $\gamma = 0.98$ a $\alpha = 0.7$
- hustota referenčného riešenia = 1/32 (4096 stavov)
- počet akcií v každom stave = 8
- hustota riedkej tabuľky = 1/8 (1:16 pomer)
- počet bázičkových funkcií $l = 64$
- rozsah parametrov

$$- \alpha_{ja}(n) \in \langle -1, 1 \rangle$$

$$- \beta_{ja}(n) \in \langle 0, 200 \rangle$$

$$- w_{ja}(n) \in \langle -4, 4 \rangle$$

$Q_{rt}(s(n), a(n))$ referenčná funkcia Q (funkcia 0), kde $t \in \langle 0, 19 \rangle$ je číslo trialu
 $Q_{jt}(s(n), a(n))$ testované funkcie Q a $j \in \langle 1, 5 \rangle$.

Celková chyba behu trialu t je

$$e_{jt} = \sum_{s,a} (Q_{rt}(s,a) - Q_{jt}(s,a))^2$$

priemerná, minimálna, maximálna chyba a smerodajná odchylka

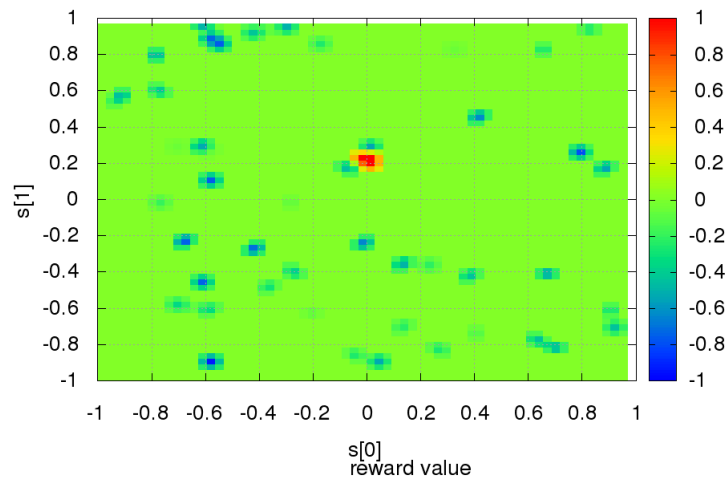
$$\bar{a}_j = \frac{1}{20} \sum_t e_{jt}$$

$$e_j^{min} = \min_t e_{jt}$$

$$e_j^{max} = \max_t e_{jt}$$

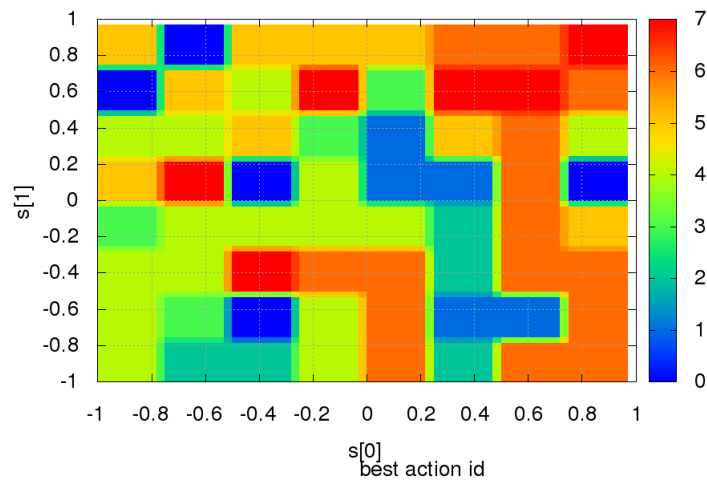
$$\sigma_j^2 = \frac{1}{20} \sum_t (\bar{a}_j - e_{jt})^2$$

4.3 Výsledky experimentu

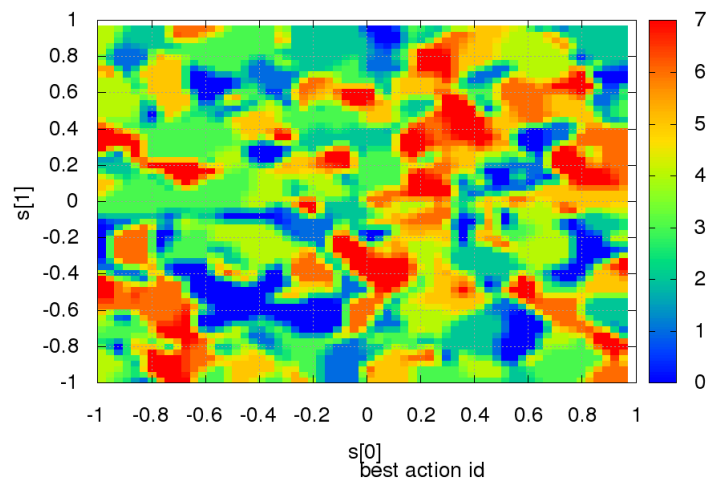


Obr. 4.2: odmeňovacia funkcia

$$e_{jt}(s) = (Q_{rt}(s,a) - Q_{jt}(s,a))^2$$



Obr. 4.3: fig:sparse table

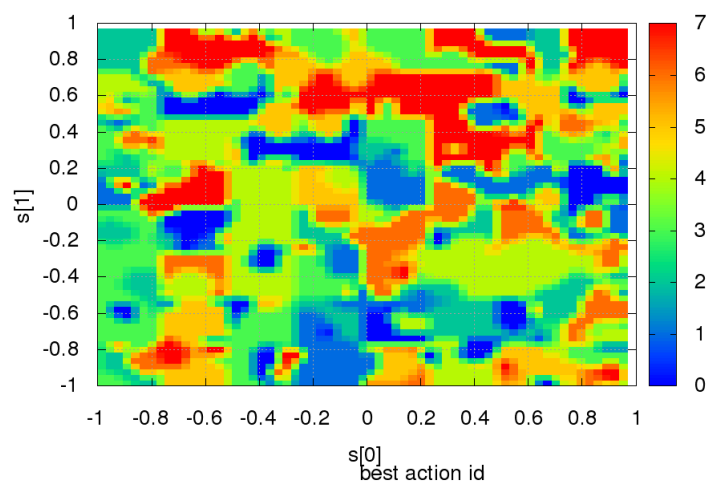


Obr. 4.4: fig:linear combination Gauss

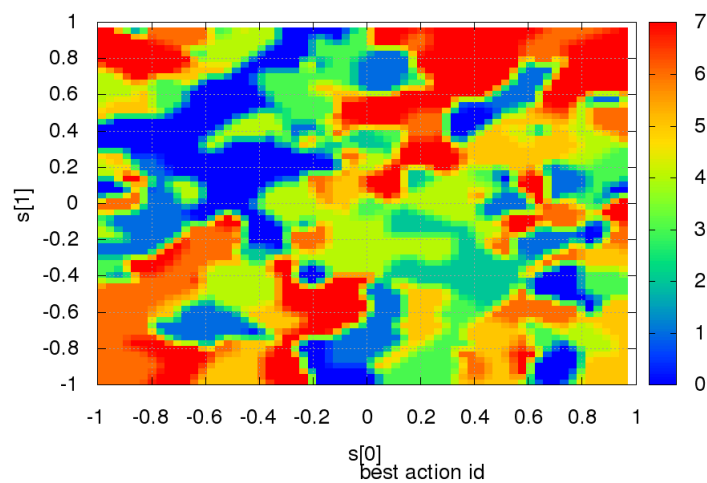
Chybové funkcie - Výsledky experimentov

$$e_{jt}(s) = (Q_{rt}(s, a) - Q_{jt}(s, a))^2$$

max $Q(s, a)$ - Výsledky experimentov



Obr. 4.5: sparse table + linear combination Gauss



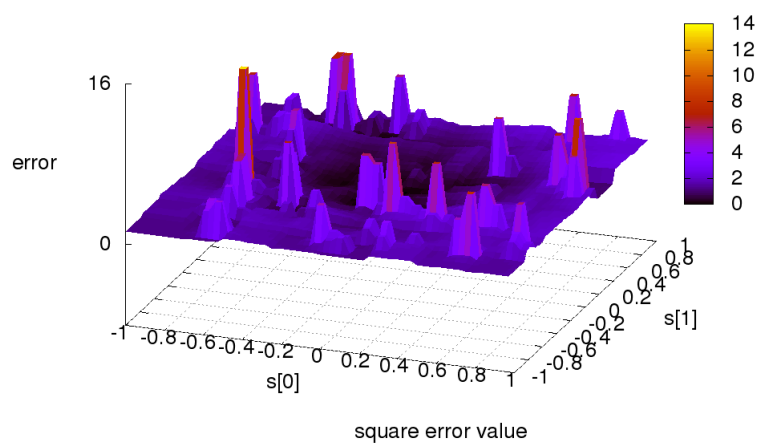
Obr. 4.6: linear combination Kohonen function

Priebeh trialov - Výsledky experimentov

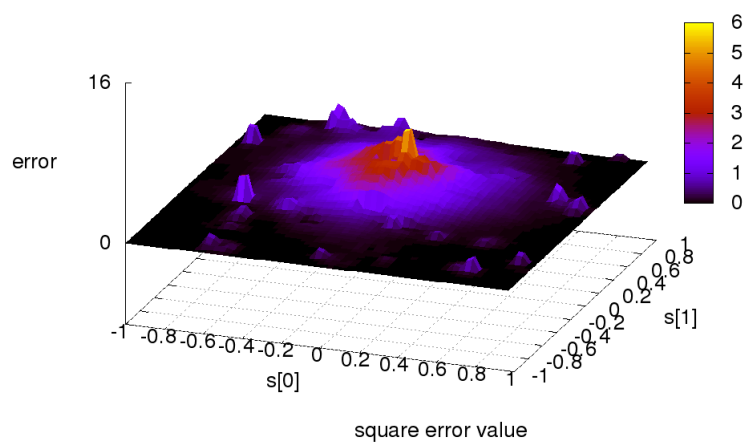
Mapa 1 - Výsledky experimentov

Mapa 0 - Výsledky experimentov

Mapa 2 - Výsledky experimentov

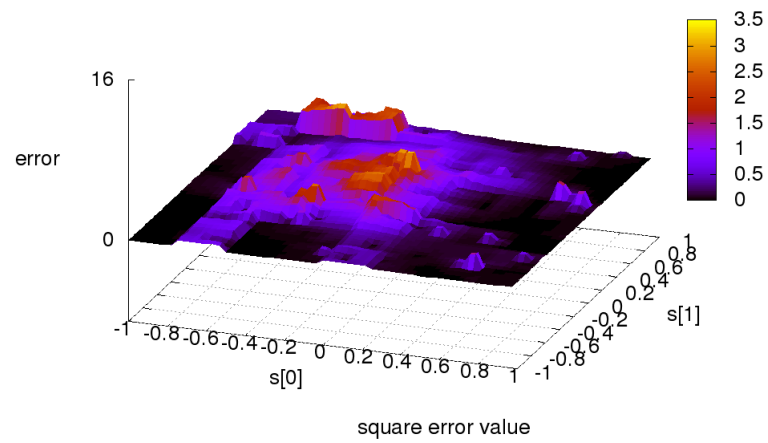


Obr. 4.7: sparse table

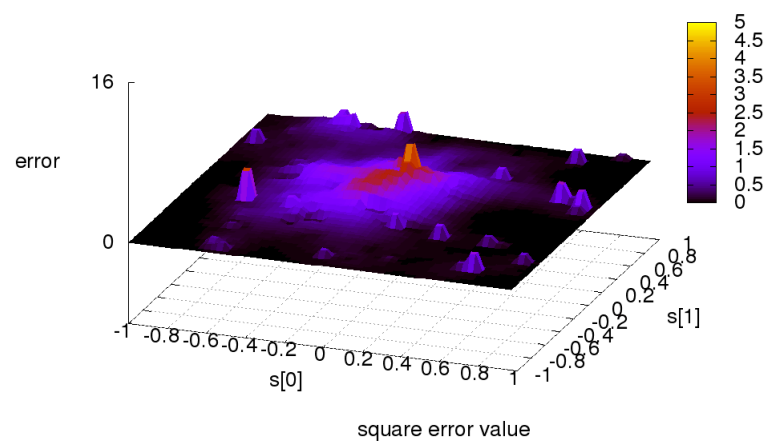


Obr. 4.8: linear combination Gauss

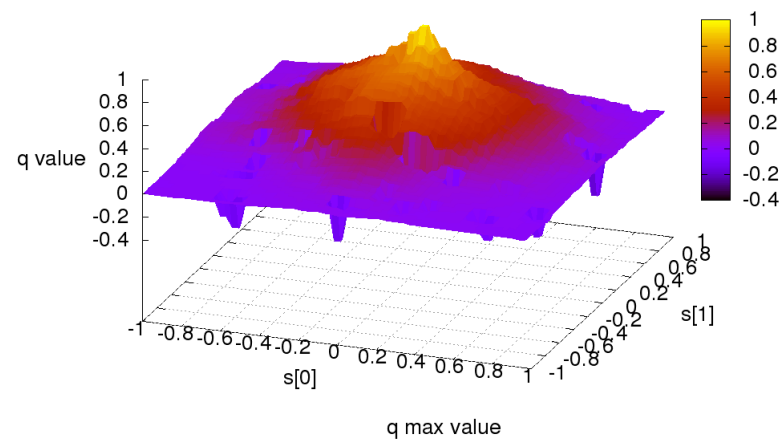
Mapa 3 - Výsledky experimentov



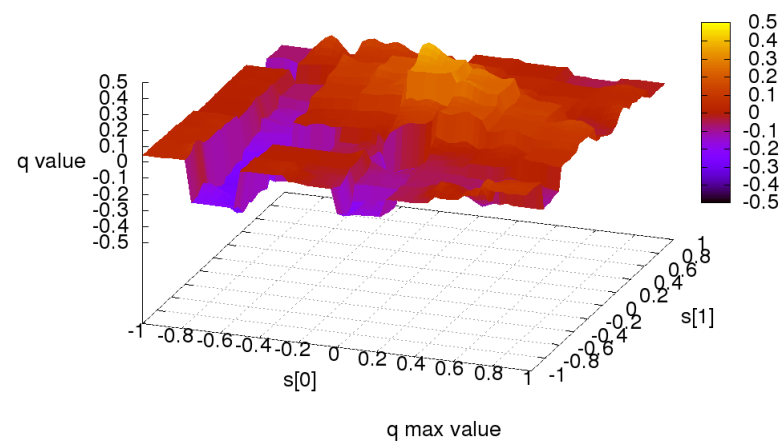
Obr. 4.9: sparse table + linear combination Gauss



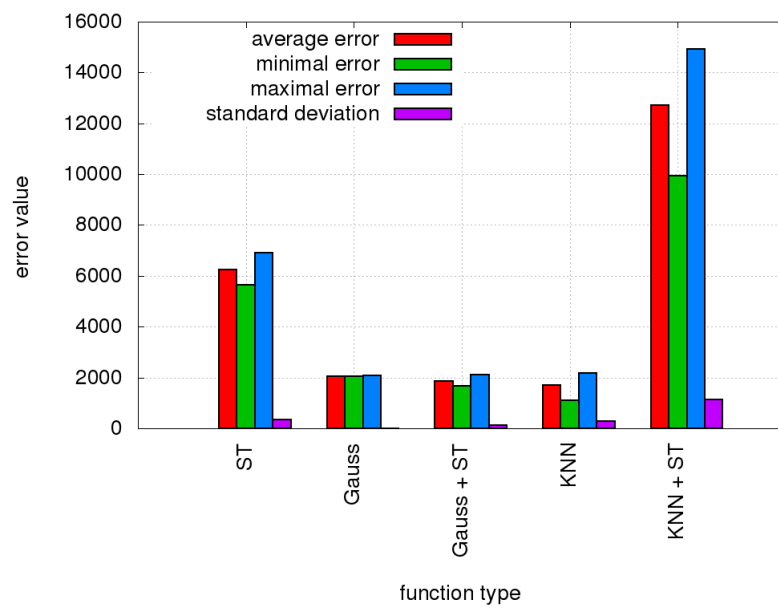
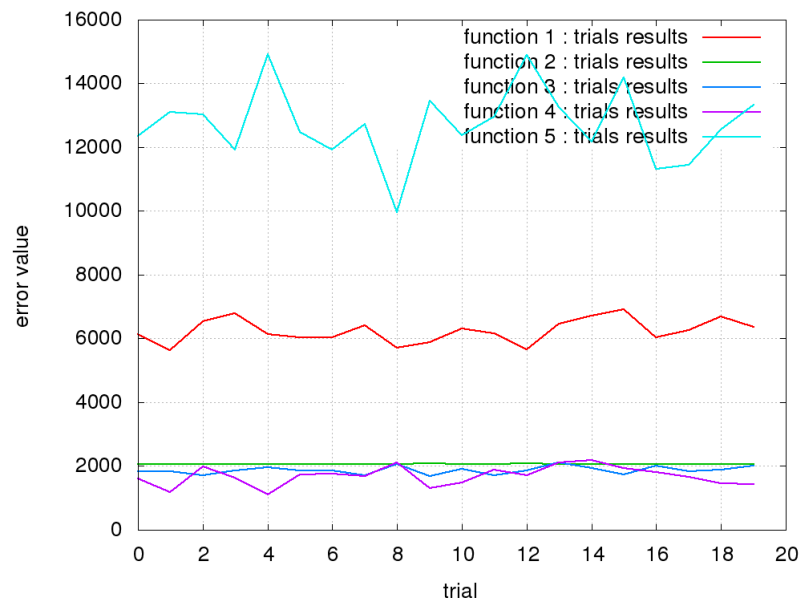
Obr. 4.10: linear combination Kohonen function

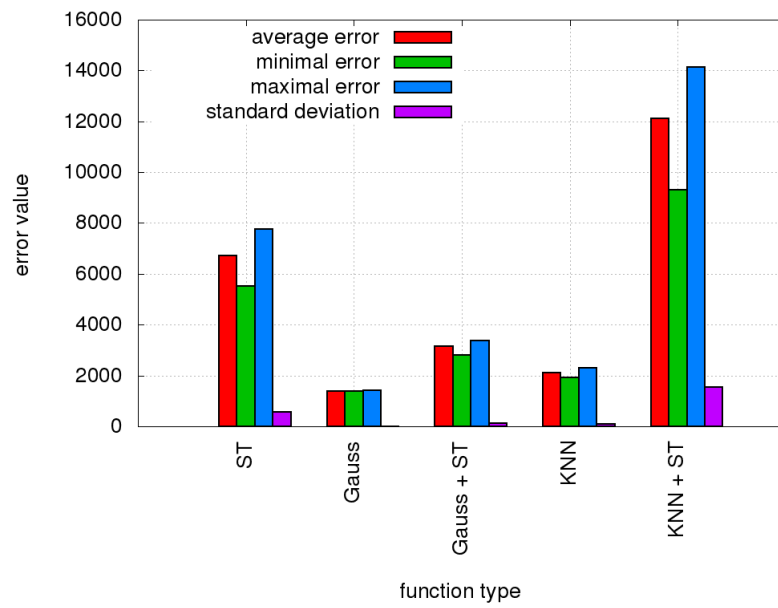
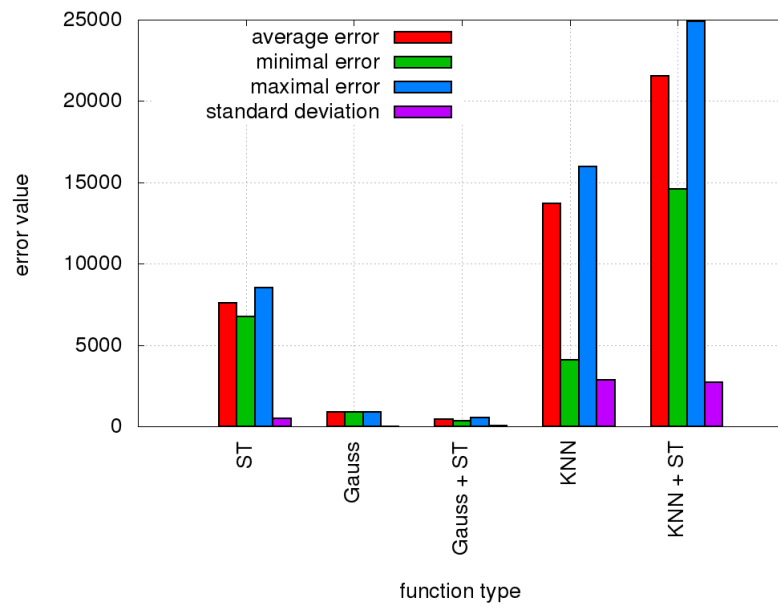


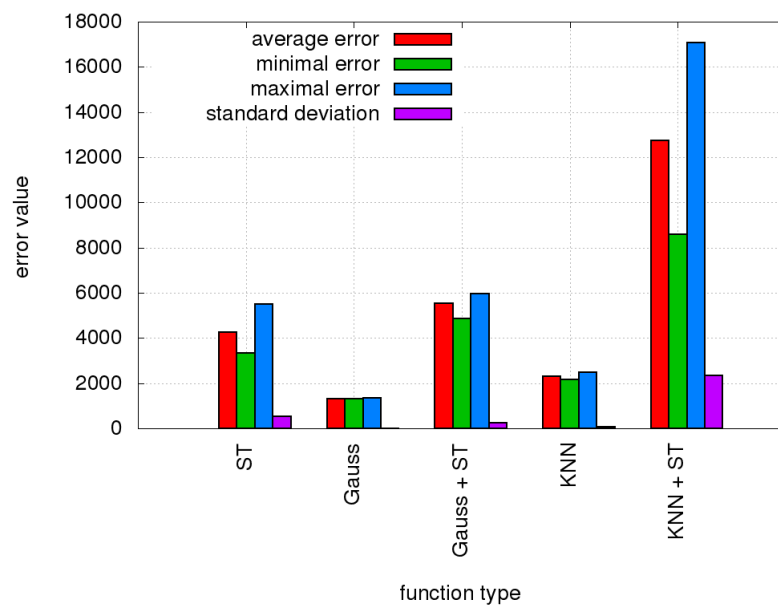
Obr. 4.11: reference table



Obr. 4.12: sparse table + linear combination Gauss







Literatúra

- [1] NanoQ learning zdrojové súbory https://github.com/michalnand/q_learning/tree/master/src/nano_q_learning
- [2] Fundamentals of Artificial Neural Networks Mohamad H. Hassoun, MIT Press, 1995
- [3] B. Irie Auditory & Visual Perception Res. Lab., ATR, Osaka, Japan, S. Miyake : Neural Networks, 1988., IEEE International Conference on, INSPEC 3350063
- [4] Kolomongorov teorém, stručne https://en.wikipedia.org/wiki/Universal_approximation_theorem
- [5] R. Rojas: Neural Networks, Springer-Verlag, Berlin, 1996, chap 7
- [6] Martin Riedmiller, Computer Standards & Interfaces Volume 16, Issue 3, July 1994, Pages 265-278 : Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms
- [7] J. Leonard, M.A. Kramer, Computers & Chemical Engineering Volume 14, Issue 3, March 1990, Pages 337–341 Improvement of the backpropagation algorithm for training neural networks
- [8] Jonathan Engel, Norman Bridge Laboratory of Physics 161-33, California Institute of Technology, Pasadena, CA 91125, USA : Teaching Feed-Forward Neural Networks by Simulated Annealing