5. prednáška

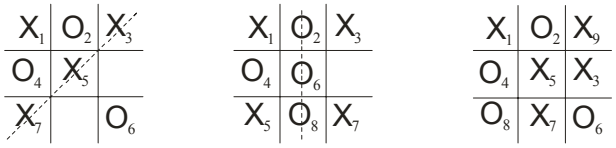
Učenie s odmenou a trestom a emergencia stratégie hry

5.1 Úvodné poznámky

Metódy riešenia zložitých úloh patria v umelej inteligencii medzi základné problémy už od jej vzniku v 50. rokoch. Prostriedkami klasickej (symbolickej) umelej inteligencie tieto metódy sa najčastejšie redukujú na rôzne prehľadávacie metódy stromov riešení, ktoré sú akcelerované rôznymi pravidlami a heuristikami. Aplikácie týchto symbolických metód na netriviálne problémy je ohraničená dimenziou ich stavových priestorov. Ak je táto dimenzia príliš veľká (N>10^5), potom prehľadávacie metódy v dôsledku existencie „kombinatoriálnej explózie poskytujú len veľmi približné výsledky v dôsledku nutnosti redukcie hĺbky prehľadávania v stromoch riešení. Preto sa v umelej inteligencii hľadajú nové prístupy a metafory k prekonaniu "kombinatoriálnej bariéry". Cieľom tohto článku je ukázať efektívnosť moderných metód umelej inteligencie, ktoré sú založené na metaforách Darwinovho prirodzeného výberu (evolučné metódy), ľudského mozgu (neurónové siete) a „učenia s odmenou a trestom“ (angl. *reinforcement learning*). Všetky tieto prístupy, samotné alebo ich kombinácia, sú v súčasnosti používané pre hľadanie riešenia zložitých úloh. Ako ilustráciu použitia týchto metód ukážeme na jednoduchej hre piškvorky (angl. *Tic-Tac-Toe*) ako zostrojiť program, ktorý sa túto hru učí hrať. Zvolený prístup k učeniu neurónovej siete bude vyžadovať , aby bol k dispozícii určitý model tejto hry. Potom na kvalite modelu závisia aj získané výsledky. Počítačový program hrá s týmto model neustále hry, pomocou učenia s odmenou a trestom hľadáme takú neurónovú sieť, ktorá je schopná hrať na výbornej úrovni proti použitému modelu.

5.2 Hra Tic-Tac-Toe (Piškvorky)

Detská hra piškvorky má dvoch hráčov, prvý hráč je označený symbolom X a druhý hráč symbolom O. Snahou každého hráča je umiestniť na štvorcovej 3×3 hracej doske svoje symboly tak, aby tvorili buď riadok, stĺpec, alebo uhlopriečku. Hra je zahájená hráčom X, potom nasleduje hráč O, toto striedanie hráčov sa opakuje tak dlho, až je vytvorená požadovaná „riadková“ pozícia alebo na hracej doske je umiestnených deväť symbolov. Hru vyhráva ten hráč, ktorý prvý vytvoril požadovaný obrazec, alebo hra končí remízou, ak hracia doska obsahuje deväť symbolov a žiaden hráč nevytvoril požadovaný obrazec (pozri obr. 5.1).



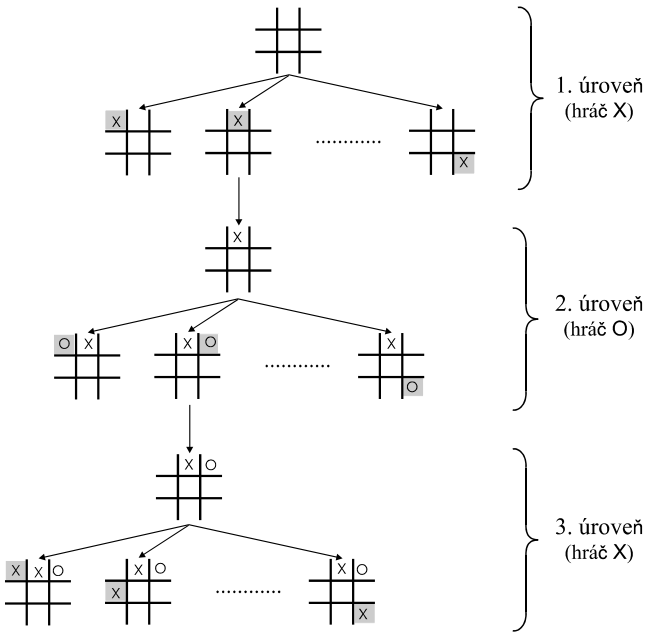
Obrázok 5.1. Znázornenie 3 partií hry piškvorky Indexy pri jednotlivých symboloch X a/alebo O znamenajú poradie ťahu. V prvých dvoch partiách bolo dosiahnuté víťazstvo, hráč umiestnil svoje symboly do "riadku" naznačeného prerušovanou čiarou. Tretia partia skončila remízou, žiadnemu hráčovi sa nepodarilo umiestniť svoje symboly do "riadku", po 9 ťahoch, keď sú obsadené všetky pozície hracej dosky, hra končí remízou.

Hra piškvorky patrí medzi tzv. symetrické hry, z pohľadu druhého hráča je hra identická s hrou prvého hráča (hráči sú rovnocenní, odlišujú sa len v tom, že jeden z nich zahajuje hru). Obaja hráči riešia rovnaký strategický problém, vytvoriť čo najrýchlejšie „riadkovú“ pozíciu svojich znakov a súčasne zabrániť súperovi vytvorenie jeho „riadkovej“ pozície.

Dimenzia stavového priestoru hry piškvorky sa zostrojiť pomocou jednoduchých kombinatorických úvah, dostaneme

$$N = \sum_{p=1}^{5} \binom{9}{p} \left[\binom{9-p}{p-1} + \binom{9-p}{p} \right] - 2 \times 6 \times 13 = 5889 \qquad (5.1)$$

kde napr. pre p=1 dostaneme 9×(1+8) pozícií, ktoré obsahujú buď len jeden symbol X, alebo dva symboly X a O. Posledný člen na pravej strane odpovedá skutočnosti, že niektoré členy, obsahujúce jeden stľpec (riadok) znakov X a jeden stĺpec (riadok) znakov O, nemôžu byť zahrnuté v povolených pozíciách, týchto "nepovolených" pozícií je 156.



Obrázok 5.2. Znázornenie stromu riešení, ktorého vrchol (koreň) je prázdna hracia doska. Po prvom ťahu, ktorý hral hráč X, je obsadená jedna pozícia (zdôraznená vytieňovaným) symbolom X. V druhej úrovni, hranej hráčom O, je obsadená pozícia (vytieňovaná) symbolom O. Vetva stromu riešení končí vtedy, ak jeden hráč zvíťazil, alebo sú obsadené všetky pozície na hracej doske - hra skončila remízou. Konštrukcia tohto stromu riešení môže byť algoritmicky jednoducho realizovaná pomocou metódy spätného prehľadávania, výsledky tejto metódy sú uvedené v predchádzajúcej tabuľke.

Horný úsek stromu riešení je znázornený na obr. 2. Koreňom tohto stromu riešení je prázdna hracia doska, v druhej úrovni je 9 pozícií, ktoré obsahujú jeden X znak. Pomocou metódy spätného prehľadávanie sme zostrojili celý strom riešení a zistili sme, že má nasledujúci počet koncových pozícií, ktoré sú charakterizované takto

No.	Počet	Typ
1	131184	víťazstvo hráča X
2	77904	víťazstvo hráča O
3	46080	remíza hráčov X a O
	255168	celkový počet

Z tejto tabuľky vyplýva, že prvý hráč X má väčšiu šancu hru vyhrať. Podrobnou analýzou sa dá ukázať, že aj hráč O môže hru forsírovať tak, že remizuje. Celkový počet koncových vetví v strome riešení možno jednoducho odhadnúť ako 9!=362880.

Pre naše ďalšie štúdie bude potrebné mať k dispozícii určitý referenčný algoritmus hry piškvorky, ktorý aj keď nie je najefektívnejší, vedie obvykle aspoň k remíze. Uvedieme jednoduchý algoritmus, ktorý sa zakladá na tom, že hráč vykoná ťah, pričom tento je určený nasledujúcimi ôsmymi pravidlami s klesajúcou prioritou

Model hry piškvorky.

1. pravidlo. *Hráč vykoná ťah, ktorý vedie k jeho víťazstvu.*

2. pravidlo. *Hráč vykoná ťah, ktorý zabráni víťazstvu oponenta v nasledujúcom ťahu.*

3. pravidlo. *Hráč vykoná ťah, ktorým si pripraví možnosť dvojitého použitia 1. pravidla* (*tzv. vidlička*).

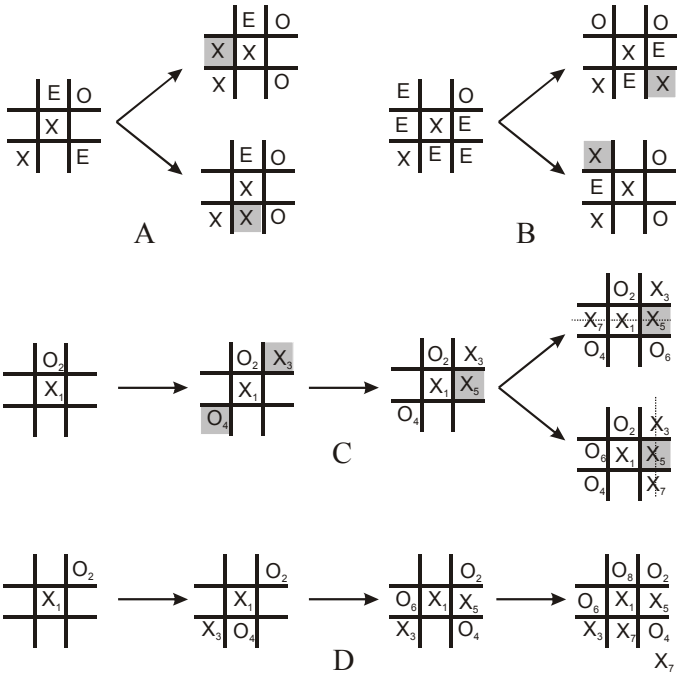
4. pravidlo. *Hráč vykoná ťah, ktorým zabráni oponentovi pripraviť "vidličku"*

5. pravidlo. *Hráč obsadí stredové pole.*

6. pravidlo. *Hráč obsadí rohové pole, ktorého proti-poloha je obsadená oponentom*

7. pravidlo. *Hráč obsadí rohové pole.*

8. pravidlo. *Hráč obsadí volné pole.*



Obrázok 5.3. Diagramy A-B znázorňujú základné typy vidličkových pozícií, ktoré sú aplikovateľné použitím pravidiel 3 a 4. Tak napríklad, diagram A znázorňuje východiskovú pozíciu pre prípravu "vydličky"; písmena E špecifikujú prázdne bunky. Z diagramu A môžeme vytvoriť dve "vydličkové" pozície. Diagram C znázorňuje hru, ktorá je prehraná pre hráča O už po druhom ťahu. Druhý hráč urobil „fatálnu“ chybu, keď umiestnil svoju figuru O v druhom ťahu v druhom stĺpci hore, podľa pravidla mal obsadiť rohové pole. Diagram D znázorňuje hru, ktorá prebieha podľa pravidiel, končí remízou. .

Model nie je plne deterministický, v prípade, že podľa niektorého aplikovaného pravidla existuje niekoľko možností, tak potom z nich vyberieme náhodne jednu z nich. Prvé dve pravidlá odpovedajú jednoduchej rekognoskácii aktuálnej pozícii o jeden ťah vopred. Rekognoskácia o dva ťahy vopred je zahrnutá až v treťom pravidle, pripravuje možnosť použitia 1. pravidla. Štvrté pravidlo, ktoré bráni oponentovi pripraviť „vidličku“, t. j. ide

o rekognoskáciu tri ťahy vopred. Na obr. 5.3, diagramy A-B, zobrazujú vybrané východzie pozície, kde je možné vytvoriť vetvenie pozícií, diagram C ilustruje priebeh hry, ktorá už po druhom ťahu druhého hráča je prehraná, prvý hráč môže vynútiť výhru.

5.3 Formalizácia hry piškvorky

V tejto kapitole sa budeme zaoberať formalizáciou hry piškvorky, ktorá je aplikovateľná pre všetky symetrické hry dvoch hráčov (šach, dáma, go, backgamon - vrhcáby, atď.). Nech aktuálne pozícia hry je popísaná premennou P, na túto pozíciu môžu byť aplikované prípustné akcie - ťahy tvoriace množinu $A(P)$. Použitím ťahu $a' \in A(P)$ pretransformujeme pozíciu P na novú pozíciu P', $P \xrightarrow{a'} P'$. Inverznú pozíciu $\overline{P}$ dostaneme z pozície P tak, že symboly X (O) zameníme za symboly O (X). Pojem inverznej pozície bude dôležitý pre formuláciu jednotného algoritmu pre symetrické hry, ktorý bude jednotný tak pre prvého ako aj druhého hráča. Použijeme multiagentový prístup, budeme predpokladať, že hru hrajú dvaja agenti G_1 a G_2, ktorý sú vybavený tzv. kognitívnym orgánom, pomocou ktorého sú schopných ohodnocovať jednotlivé nasledujúce pozície. Formulácia algoritmu je nasledovná (pozri obr. 5):

Algoritmus 0.

Krok 1. Hra je zahájená prvým hráčom, $G \leftarrow G_1$, a počiatočnou pozíciou, $P \leftarrow P_{\text{ini}}$.

Krok 2. *Hráč G vytvorí z pozície P množinu nasledujúcich pozícií* $A(P) = \{P_1, P_2, \ldots, P_n\}$.

(a) Ak množina je prázdna, potom oba hráči G_1 a G_2 remizujú a hra pokračuje krokom 4.

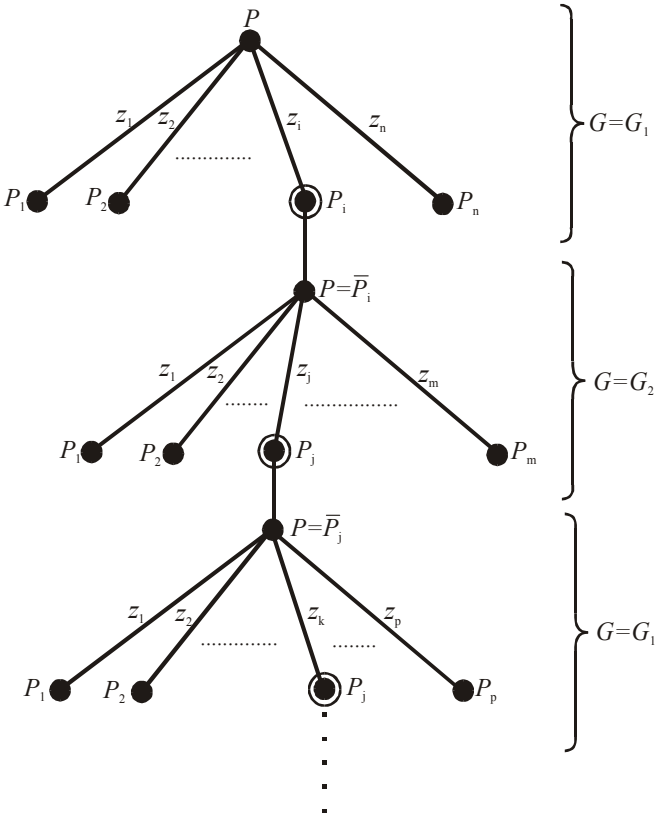
(b) V prípade, že hráč je prvý a hrá sa prvý ťah, tak náhodne vyberieme jednu pozíciu z $A(P)$. V opačnom prípade každá pozícia P_i z množiny nasledujúcich pozícií je ohodnotená koeficientom $0<z_i<1$. Hráč vyberie za nasledujúcu pozíciu takú $P' \in A(P)$, ktorá je ohodnotená maximálnym koeficientom z, $P \leftarrow P'$. Ak pozícia P je víťazná, potom hráč G víťazí a hra pokračuje krokom 4.

Krok 3. *Hra prechádza na druhého hráča, $G \leftarrow G_2$, pozíciu P si vytvorí inverziou aktuálnej pozície, $P \leftarrow \overline{P}$, hra pokračuje krokom 2.*

Krok 4. *Koniec hry.*

Podobne, ako pre model hry piškvorky (pozri 2. kapitolu), tento algoritmus nie je plne deterministický, v prípade prvého hráča a jeho prvého ťahu, ťah sa vyberie z množiny možných ťahov náhodne, potom je už hra plne deterministická. Kľúčovú úlohu v algoritme má výpočet koeficientov $z=z(P')$ pre pozície $P' \in A(P)$. Toto môže byť vykonané buď metódami klasickej umelej inteligencie, ktoré sú založené na kombinácií metód spätného prehľadávania a rôznych heuristík. Našu pozornosť usmerníme na moderný prístup multiagentových systémov, ktorý sa zakladá na predpoklade, že správanie sa agenta v prostredí a/alebo vykonávanie určitých činností je plne determinované jeho kognitívnym orgánom, ktorý vykazuje určitú plasticitu (t.j. je schopný učenia). Kognitívny orgán agenta G je obykle numericky realizovaný parametrickým zobrazením $G(w):P \rightarrow R$, kde w je parameter (parametre) zobrazenia a ktoré priradzuje každej pozícii (prostrediu) reálne číslo z, formálne $z=G(P,w)$. Zmenou w menia sa vlastnosti kognitívneho orgánu (t.j. dostávame iné

parametrické zobrazenie), týmto spôsobom máme zabezpečenú plasticitu kognitívneho orgánu, čo je nutná podmienka procesu učenia agenta v danom prostredí, v procesu ktorého stále lepšie a lepšie zvláda požadovanú úlohu. Vo všeobecnosti parametrické zobrazenie $G(w):P\rightarrow R$ môže byť realizované mnohými rôznymi spôsobmi, od ktorých sa požaduje len splnenie podmienky „univerzálneho aproximátora“. Táto všeobecná podmienka je automaticky splnená v rámci požadovanej presnosti pre mocninné rozvoje alebo pre neurónové siete s dopredným šírením signálu a aspoň s jednou vrstvou skrytých neurónov.



Obrázok 5.5. Diagramatická reprezentácia symetrickej hry pre dvoch hráčov. Vybraný hráč G, z aktuálnej pozície vytvorí všetky možné nasledujúce pozície P_1, P_2, … . Použitím svojho kognitívneho orgánu (reprezentovaného napríklad neurónovou sieťou) ohodnotí každú novú pozíciu číslom λ, ako svoj nasledujúci ťah vyberie si tú pozíciu, ktorá má maximálne ohodnotenie. Tento elementárny akt rozhodovania sa opakuje tak, že hráči sa striedajú. Oba hráči majú rovnocenný kognitívny orgán, to znamená, že ich deklarácia na prvého a druhého hráča je náhodná. Z tejto skutočnosti vyplýva, že druhý hráč používa svoj kognitívny orgán na invezné pozície $\overline{P}$, pristupuje k svojej aktuálnej pozícii ako keby bol prvý hráč. V opačnom prípade, keby hráč nerozlišoval či je prvý alebo druhý hráč, musel by mať dva kognitívne orgány, ktoré by používal v závislosti od toho, či je prvý alebo druhý hráč.

Budeme predpokladať, že agent je schopný svojim kognitívnym orgánom ohodnocovať vznikajúce nasledujúce pozície tak v prípade, že je prvý, alebo aj druhý hráč. Tento predpoklad je plazibilný pre symetrické hry, kde strategické pravidlá tak pre prvého, ako aj druhého hráča sú rovnaké, určitý malý rozdiel však existuje v tom, že hru zahajuje prvý hráč a tak môže v určitom rozsahu vnucovať štýl hry druhému hráčovi. Vo vyššie uvedenom algoritme je táto požiadavka univerzálnosti kognitívneho orgánu (t.j. agent nemá dva kognitívne orgány, jeden pre prípad, že je prvým hráčom a iný pre prípad, že je druhým hráčom) realizovaná tým, že druhý hráč svojim kognitívnym orgánom neohodnocuje priamo pozície $P'\in A(P)$ ale ich inverzie $\overline{P}'$, t.j. agent aj ako druhý hráč ohodnocuje nasledujúce

pozície z pohľadu prvého hráča, čiže kognitívny orgán je nastavený na to, že agent je akoby vždy prvým hráčom, v prípade, že je druhým hráčom, „ohodnocuje“ inverzné pozície.

5.4 Štruktúra kognitívneho orgánu - neurónovej siete

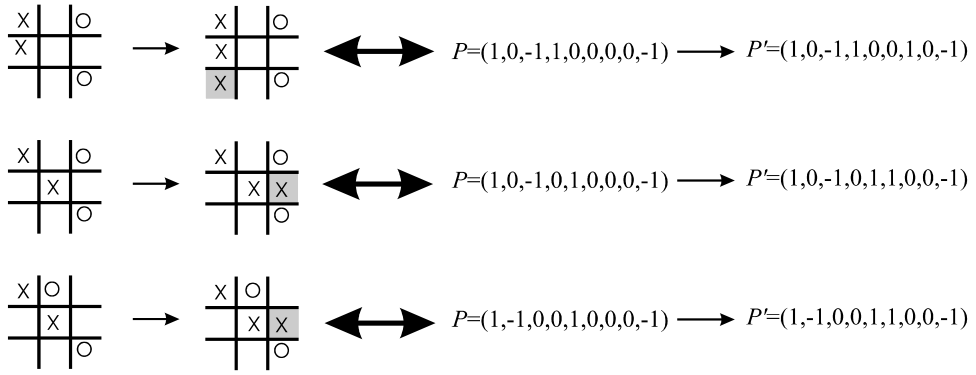
Prv ako pristúpime k špecifikácii kognitívneho orgánu agentov, musíme zaviesť tzv. numerickú reprezentáciu pozícií. Pozícia je reprezentovaná 9-rozmerným vektorom

$$x(P) = (x_1, x_2, \ldots, x_9) \in \{0,1,-1\}^9 \tag{5.2a}$$

kde jednotlivé zložky určujú jednotlivé políčka v pozícii P

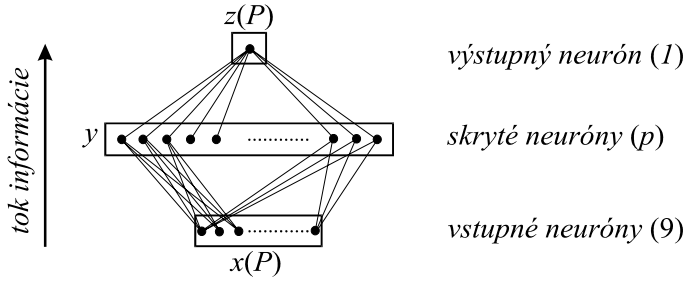
$$x_i = \begin{cases} 0 & (i-\text{té pole je neobsadené}) \\ 1 & (i-\text{té pole je obsadené X}) \\ -1 & (i-\text{té pole je obsadené O}) \end{cases} \tag{5.2b}$$

Tak napríklad, numerická reprezentácia pozícií na obr. 3 je znázornená na obr. 6.



Obrázok 5.6. Pozície sú reprezentované 9-rozmerným vektorom z $\{-1,0,1\}^9$. Ak je políčko v pozícii obsadené symbolom X (O) , potom príslušná zložka vektora sa rovná 1 (-1). V prípade, že políčko je prázdne, potom zložka vektora je nulová.

Použitá neurónová sieť má architektúru typu „feed-forward“, s jednou vrstvou skrytých neurónov. Aktivity vstupných neurónov sú určené numerickou reprezentáciou $x(P)$ danej pozície P, výstupná aktivita sa rovná ohodnoteniu pozície $\lambda(P)$ (pozri obr. 7), počet parametrov neurónovej siete je $11p+1$, kde p je počet skrytých neurónov. Plasticita takto špecifikovaného kognitívneho aparátu pomocou neurónovej siete je je špecifikovaná dvoma spôsobmi: (a) š**trukturálna plasticita,** ktorá znamená, že sa mení počet skrytých neurónov (t.j. mení sa štruktúra – topológia neurónovej siete), a (b) **parametrická plasticita**, ktorá odpovedá zmene váh spojov a prahových koeficientov skrytých a výstupných neurónov.



Obrázok 5.7. Dopredná (feed forward) neurónová sieť s jednou vrstvou skrytých neurónov. Vstupné aktivity sú rovné 9-rozmernému vektoru $\boldsymbol{x}(P)$, ktorý kóduje pozície hry. Výstupná aktivita sa rovná reálnemu číslu $z(P)$ z otvoreného intervalu (0,1), toto číslo je ohodnotenie vstupnej pozície.

Aktivity neurónov skrytých neurónov a výstupného neurónu sú určené vzťahmi

$$y_i = t\left(\sum_{j=1}^{9} w_{ij} x_j + \vartheta_i\right) \quad (i = 1,2,\ldots,p) \tag{5.3a}$$

$$z = t\left(\sum_{i=1}^{p} \tilde{w}_j y_j + \tilde{\vartheta}\right) \tag{5.3b}$$

kde $t(\xi)$ je prechodová funkcia tvaru jednoduchej sigmoidy (pozri obr.5.8)

$$t(\xi) = \frac{1}{1+e^{-\xi}}, \qquad t: R \to (0,1) \quad \Rightarrow \quad 0 < \lambda < 1 \tag{5.4}$$

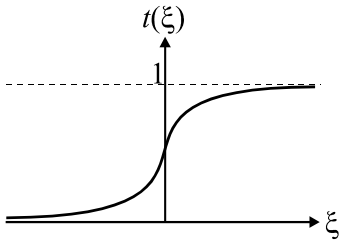
Pri adaptácii neurónových sietí sa obvykle využíva parciálne derivácie výstupnej aktivity $\lambda(P)$ vzhľadom k parametrom neurónovej siete. Tieto parciálne derivácie vzhľadom k parametrom výstupného neurónu sú určené takto

$$\frac{\partial z(P)}{\partial \tilde{\vartheta}} = z(P)\left[1 - z(P)\right], \quad \frac{\partial z(P)}{\partial \tilde{w}_j} = \frac{\partial z(P)}{\partial \tilde{\vartheta}} y_j \tag{5.5a}$$

Podobným spôsobom sú určené aj parciálne derivácie vzhľadom k parametrom skrytých neurónov

$$\frac{\partial z(P)}{\partial \vartheta_i} = y_i(1-y_i)\frac{\partial z(P)}{\partial \tilde{\vartheta}} \tilde{w}_i, \quad \frac{\partial z(P)}{\partial w_{ij}} = \frac{\partial z(P)}{\partial \vartheta_i} x_j \tag{5.5b}$$

Parciálne derivácie môžeme jednoducho počítať rekurentným postupom nazývaný „metóda spätného šírenia“, najprv sa spočítajú parciálne derivácie pre parametre výstupného neurónu, potom môžeme spočítať parciálne derivácie vzhľadom k parametrom skrytých neurónov.



Obrázok 5.8. Priebeh funkcie sigmoidy definovanej vzťahom (4). Táto funkcia je kladná, monotónne rastúca a vyhovujúca asymptotickým podmienkam $t(\xi)\to 1$, pre $\xi\to\infty$ a $t(\xi)\to 0$, pre $\xi\to -\infty$.

5.5 Klasický prístup k adaptácii neurónových sietí reprezentujúcich kognitívny orgán agenta

Klasický prístup k adaptácii neurónových sietí je založený na použití tzv. tréningovej množiny, ktorá obsahuje dvojice vektorov vstupných a požadovaných výstupných aktivít (poznamenajme, že vektor vstupných aktivít obvykle je numerickou reprezentáciou objektu

klasifikovaného neurónovou sieťou, pričom požadované výstupné aktivity klasifikujú daný objekt)

$$\mathrm{A}_{train} = \left\{ \boldsymbol{x} = \left(x_1, x_2, ..., x_n \right) / z_{req} \right\} \tag{5.6}$$

To znamená, že neurónová sieť je parametrické zobrazenie *n*-rozmerných vektorov na reálne číslo z otvoreného intervalu (0,1)

$$G\left(w\right): R^n \to \left(0,1\right) \tag{5.7}$$

Adaptácie neurónovej siete je realizovaná pomocou *minimalizácie* účelovej funkcie vyjadrujúcej sumu kvadrátov rozdielov medzi vypočítanými a požadovanými výstupnými aktivitami

$$E\left(w\right) = \frac{1}{2} \sum_{\boldsymbol{x}/z_{req} \in A_{train}} \left(z_{req} - G\left(\boldsymbol{x}; w\right) \right)^2 \tag{5.8}$$

Váhové koeficienty neurónovej siete sú upravované pomocou gradientovej metódy najprudšieho spádu (steepest descent)

$$w := w - \alpha \frac{\partial E}{\partial w} = w + \Delta w \tag{5.9a}$$

$$\Delta w = \alpha \sum_{x/z_{req} \in A_{train}} \left(z_{req} - G\left(\boldsymbol{x}, w\right) \right) \frac{\partial G\left(\boldsymbol{x}, w\right)}{\partial w} \tag{5.9b}$$

kde α je kladný parameter nazývaný „krok učenia“. Výpočet parciálnych derivácií výstupných aktivít vzhľadom k parametrom neurónovej siete je popísaný vzťahmi (5.5a) a (5.5b).

Po všeobecnom úvode o štandardnom spôsobe adaptácie neurónových sietí obrátime našu pozornosť o použití tohto typu adaptácie neurónových sietí, ktoré slúžia ako kognitívny orgán aganta hrajúceho symetrickú hru pre dvoch hráčov. Pre väčšiu konkrétnosť našich úvah budeme konkrétne vždy hovoriť o piškvorky hre, avšak naše úvahy budú dostatočne všeobecné, aby boli aplikovateľné aj pre iné typy symetrických hier. Množinu P všetkých možných pozícií rozdelíme na dve disjunktné podmnožiny, ktoré obsahujú párny a nepárny počet znakov

$$\mathrm{P} = \mathrm{P}_{odd} \cup \mathrm{P}_{even} \tag{5.7a}$$

$$P_{odd} = \left\{ \left(x_1, x_2, ..., x_9 \right); \sum_{i=1}^{9} \left| x_i \right| = 2k+1, k = 0,1,...,4 \right\} \tag{5.7b}$$

$$P_{even} = \left\{ \left(x_1, x_2, ..., x_9 \right); \sum_{i=1}^{9} \left| x_i \right| = 2k, k = 1,...,4 \right\} \tag{5.7c}$$

Pozície z P_{odd}, obsahujúce nepárny počet znakov, sa vyskytujú v priebehu hry piškvorky u prvého hráča, zatiaľ čo, pozície z $\mathrm{P}_{\mathrm{even}}$, obsahujúce párny počet znakov, sa vyskytujú v priebehu hry piškvorky u druhého hráča. Potom z pohľadu toho-ktorého hráča, pozície z príslušnej tréningovej podmnožiny môžu byť ohodnotené reálnym číslom z_{req}, ktoré špecifikuje či daná pozícia vedie k víťaztvu (z_{req}=1), remíze (z_{req}=1/2) a prehre (z_{req}=0). Pre lepšie pochopenie týchto myšlienok obrátime našu pozornosť na obr. 5.4, kde je uvedená pozícia *P*=(0,-1,0,0,1,0,0,0,0), ktorá pri optimálnej hre prvého hráča vedie k jeho víťazstvu (táto skutočnosť je nezávislá od hry druhého hráča). To znamená, že pozícia *P*=(0,-1,0,0,1,0,0,0,0)$\in$ $\mathrm{P}_{\mathrm{even}}$ je ohodnotená z_{req}=1. Vo všeobecnosti, hodnotenie pozícií je netriviálna záležitosť a môže byť realizovaná s ohraničenou presnosťou "expertom", ktorý na základe svojich vedomostí, skúseností a intuície ohodnotí každú pozíciu z tréningovej množiny číslom z_{req}. Pre hru piškvorky, ktorá obsahuje „len“ okolo 8000 prípustných pozícií, je možné ohodnotenie pozícií vykonať pomocou „presnej“ metódy spätného prehľadávania. Žiaľ, tento priamočiary prístup je nerealizovateľný pre hry s podstatne väčším stavovým

priestorom (napr. šach, dáma,…), časová náročnosť metódy spätného prehľadávania prudko rastie (exponenciálne) s počtom možných pozícií.

Predpokladajme, že pozície z množiny P máme ohodnotené číslami $z_{\text{req}} \in [0,1]$, potom môžeme zostrojiť tréningovú množinu takto

$$A_{train} = \left\{ \boldsymbol{x} / z_{req} \right\} \tag{5.8}$$

to znamená, že každá pozícia z množiny P je ohodnotená reálnym číslom $z_{\text{req}} \in [0,1]$ v závislosti od toho, či pozícia pri optimálnej hre daného hráča vedie buď k víťazstvu, remíze, alebo prehre. Tréningová množina (8) môže byť použitá ako základ pre realizáciu adaptačného procesu neurónovej siete – kognitívneho orgánu, ktorého parametre v priebehu učenia sa nastavia tak, aby kognitívny orgán správne klasifikoval pozície, t.j. predpokladá sa, že výstupná aktivita neurónovej siete aproximuje čísla *z*

$$z_{req} \approx G\left(\boldsymbol{x}, w\right) \qquad \forall \, \boldsymbol{x} / z_{req} \in A_{train} \tag{5.9}$$

Obrazne povedané, úspešne adaptovaný kognitívny orgán klasifikuje pozície podobne, ako „expert“, pomocou ktorého sme vytvorili tréningovú množinu.

Niekoľko záverečných poznámok k tomuto priamočiaremu použitiu neurónovej siete k implemenmtácii algoritmu pre symetrickú hru dvoch hráčov. Aj keď je koncepčne veľmi jednoduchý a jedná sa o bezprostrednú aplikáciu neurónových sietí k problematike počítačovej implementácie symetrických hier, obsahuje vážnu reštrikciu ako je konštrukcia tréningovej množiny pre účely adaptačného procesu. Ohodnotenie každej pozície číslom λ pomocou „experta“ je realizovateľné s dostatočnou presnosťou len pre hry s relatívne malým stavovým priestorom. Podobne, použitie systematickej metódy „spätného prehľadávania“ je aplikovateľné podobne len k hrám s malým stavovým priestorom, časová náročnosť „spätného prehľadávania“ prudko rastie so zväčšovaním stavového priestoru. Naznačený prístup bol úspešne použitý pre hru backgammon Geraldom Tesaurom z IBM v r. 1989 v programe *Neurogammon*, kde rozsiahla tréningová množina bola vytvorená analýzou mnoho tisíc „majstrovských partií“. Vytvorený program používajúci neurónovú sieť na ohodnotenie pozícií tejto hry patrila svojho času k najlepším programom pre hru backgammon a v r. 1989 vyhral olympiádu programov pre túto hru.

5.6 Adaptácia kognitívneho orgánu agenta pomocou učenia s odmenou a trestom

Obrázok 5.9. E. L. Thorndike (1887-1949)

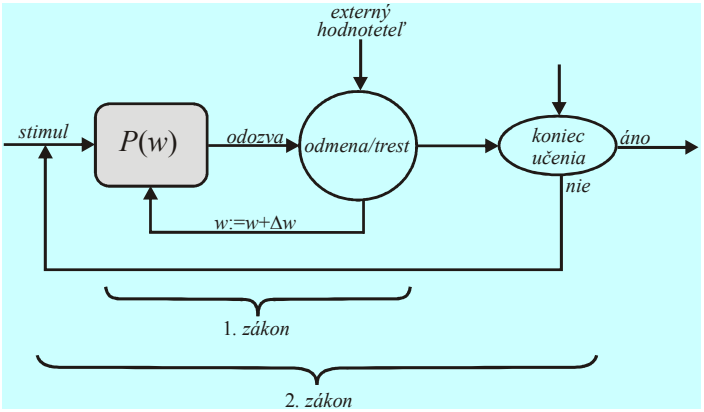
Základné idey učenia s odmenou a trestom boli naformulované americkým psychológ Edward Lee Thorndike (1887-1949), ktorý v knihe vo svojej knihe „*The Fundamentals of Learning*“ zaviedol dva zákony:

1. Zákon účinku:

Ak odozva na opakujúci sa stimul je kladná (odmena), potom väzba medzi stimulom a odozvou sa postupne zosilňuje. V opačnom prípade, ak odozva je záporná (trest), potom väzba medzi stimulom a odozvou postupne zaniká.

2. Zákon opakovaného používania:
Požadované správanie je výsledkom častého používania dvojica stimul a odozva

Učenie, ktoré je založené na týchto dvoch zákonoch sa nazýva „**učenie s odmenou a trestom**“ („**reinforcement learning**“). Tvorí teoretický základ behavioristického prístupu k učeniu (pozri obr. 5.10).



Obrázok 5.10. Schematické znázornenie učenia s odmenou a trestom.

V tejto kapitole uvedieme základné princípy modernej metódy učenia s odmenou a trestom, ktorá v súčasnosti patrí medzí účinné algoritmické prostriedky adaptácie kognitívnych orgánov multiagentových systémov. Základné princípy tohto učenia sú tieto: Agent sleduje závislosť medzi vstupným obrazcom a výstupným signálom jeho kognitívneho orgánu (ktorý sa často nazýva „akcia“ alebo „riadiaci signál“). Na základe externého skalárneho signálu „odmeny“ (reward) vyhodnocuje kvalitu výstupného signálu. Cieľom učenia je taká modifikácia kognitívneho orgánu agenta, aby výstupné signály maximalizovali príjem externých „reward“ signálov. V mnohých prípadoch signál „odmeny“ je časovo oneskorený, prichádza po dlhom slede akcií až na záver a môže byť chápaný ako ohodnotenie celej sekvencie akcií, toho či viedla k požadovanému výsledku alebo nie. V tomto prípade agent musí riešiť tzv. problém „temporal difference“ priradenia, kde učenie je založené na diferenciách medzi dočasne vykonaných predikciách pre jednotlivé elementy celkovej sekvencie akcií.

V tejto kapitole naznačíme konštrukciu metódy učenia s odmenou a trestom, ako určité zovšeobecnenie štandardnej metódy učenia neurónových sietí popísanej v predchádzajúcej kapitole. Predpokladajme, že poznáme sekvenciu pozícií a jej ohodnotenie reálnym číslom, ktoré odpovedajú pozíciám daného agenta – hráča, ktoré musel ohodnocovať číslom z

$$P_1, P_2, \ldots, P_m, z_{reward} \tag{5.10}$$

kde z_{reward} je vonkajšie ohodnotenie sekvencie postupnosti a odpovedá skutočnosti či posledná pozícia P_m je pre daného agenta – hráča víťazná, remízová, alebo prehratá.

$$z_{reward} = \begin{cases} 1 & (sekvencia\ pozícií\ je\ vítazná) \\ 0.5 & (sekvencia\ pozícií\ je\ remízová) \\ 0 & (sekvencia\ pozícií\ je\ prehraná) \end{cases} \tag{5.11}$$

Zo sekvencie (10) vytvoríme m dvojíc pozícií a ich ohodnotenie číslom z_{reward}, vytvoríme nasledujúcu účelovú funkciu

$$E(w) = \frac{1}{2}\sum_{t=1}^{m}\left(z_{reward} - G(\boldsymbol{x}_t; w)\right)^2 \tag{5.12}$$

Budeme hľadať také váhové koeficienty neurónovej siete - kognitívneho orgánu, ktoré minimalizujú účelovú funkciu. v prípade, že sa nám podarí nájsť také váhové koeficienty siete, pre ktoré je účelová funkcia nulová, potom každá pozícia zo sekvencie (10) je ohodnotená číslom z_{reward}. Rekurnetná formula pre obnovu váhových koeficientov má tvar (pozri (5.9a) a (5.9b))

$$w := w - \alpha \frac{\partial E}{\partial w} = w + \Delta w \tag{5.13a}$$

$$\Delta w = \alpha \sum_{t=1}^{m} \left(z_{reward} - z_t \right) \frac{\partial z_t}{\partial w} \tag{5.13b}$$

kde $z_t = G(P_i, w)$ je ohodnotenie t-tej pozícii P_t pomocou neurónovej sieti - kognitívnejo orgánu číslom z_t. Našim cieľom bude, aby všetky pozície zo sekvencie (5.10) boli ohodnotené rovnakým číslom z_{reward}, ktoré nám špecifikuje, či hra daného hráča pozostávajúca zo sekvencií (10) bola víťazná, remízová, alebo viedla k prehre. Výraz v zátvorke v (5.13b) dá sa jednoducho prepísať do nasledujúceho tvar.

$$\begin{aligned} z_{reward} - z_t &= z_{m+1} - z_t \\ &= z_{m+1} - z_m + z_m - z_t \\ &= z_{m+1} - z_m + z_m - z_{m-1} + z_{m-1} - P_t \\ &= \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &= \sum_{k=t}^{m} \left(z_{k+1} - z_k \right) \end{aligned} \tag{5.14}$$

K prepisu formule (13b) použijeme vzťah kde $z_{m+1} = z_{\text{rewar.d}}$. Dosadením vzťahu (5.14) do (5.13b) a použitím jednoduchej algebraickej identity

$$\sum_{t=1}^{m} \sum_{k=t}^{m} A_{kt} = \sum_{t=1}^{m} \sum_{k=1}^{t} A_{tk} \tag{5.15}$$

dostaneme

$$\begin{aligned} \Delta w &= \alpha \sum_{t=1}^{m} \left(\sum_{k=t}^{m} \left(z_{k+1} - z_k \right) \right) \frac{\partial z_t}{\partial w} \\ &= \alpha \sum_{t=1}^{m} \left(z_{t+1} - z_t \right) \sum_{k=1}^{t} \frac{\partial z_t}{\partial w} \end{aligned} \tag{5.16}$$

Sumarizujúc, inkrement obnovovacej formule váhového koeficienta má tvar

$$\Delta w = \sum_{t=1}^{m} \Delta w_t \tag{5.17a}$$

$$\Delta w_t = \alpha \left(P_{t+1} - P_t \right) \sum_{k=1}^{t} \frac{\partial z_k}{\partial w} \tag{5.17b}$$

Tento dôležitý výsledok môže byť "zovšeobecnený" na formulu, ktorá tvorí základ TD(λ) rodiny učiacich metód

$$\Delta w_t = \alpha \left(z_{t+1} - z_t \right) \sum_{k=1}^{t} \lambda^{t-k} \frac{\partial z_k}{\partial w} \tag{5.18}$$

kde parameter $0 \le \lambda \le 1$. Pre $\lambda = 1$ formula (5.18) poskytuje pôvodný výsledok (5.16), zatiaľ čo pre $\lambda = 0$ dáva

$$\Delta w_t = \alpha \left(z_{t+1} - z_t \right) \frac{\partial z_t}{\partial w} \tag{5.19}$$

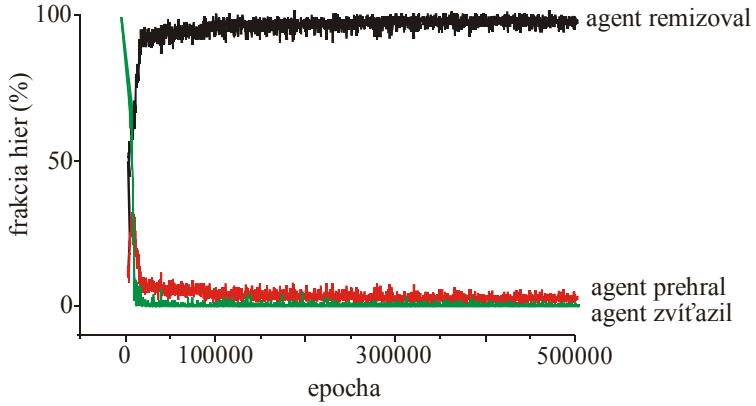
t.j. inkrement váhového koeficienta je určený len posledným "pozorovaním".

Formule (5.17) a (5.18) umožňujú rekurentý výpočet inkrementu Δw. Zavedieme nový symbol $e_t(\lambda)$, ktorý sa dá jednoducho počítať rekurentne

$$e_t(\lambda) = \sum_{k=1}^{t} \lambda^{t-k} \frac{\partial z_k}{\partial w} \Rightarrow e_{t+1}(\lambda) = \lambda e_t(\lambda) + \frac{\partial z_{t+1}}{\partial w} \tag{5.20}$$

kde $e_1(\lambda) = \partial z_1 / \partial w$. Potom jednotlivé "parciálne" inkrementy Δw_t sú určené takto

$$\Delta w_t = \alpha(z_{t+1} - z_t) e_t(\lambda) \tag{5.21}$$



Obrázok 5.11. Priebeh frakcií hier (zo 100), ktoré hral agent proti modelu hry, pričom polovicu hier (50) hral ako prvý a druhú polovicu hier hral ako druhý. Z priebehu jednotlivých prípadov jasne vyplýva, že neurónová sieť je schopná tak kvalitnej spontánnej adaptácie, že dokáže neprehrávať s modelom.

5.7 Adaptácia kognitívneho orgánu agenta pomocou učenia s odmenou a trestom

V tejto kapitole budeme študovať prvý ilustratívny príklad, kde agent hrá proti modelu hry piškvorky, popísaného v kapitole 5.2. Pripomeňme, že sa jedná o pomerne kvalitný model, ktorý je schopný „predvídania dva ťahy vopred, menovite je schopný plánovania „vidličkových" pozícií, proti ktorým protihráč, ak ich pripustí, je bezmocný. Ilustratívny príklad je realizovaný pomocou nasledujúceho algoritmu:

Algoritmus 1.

Krok 1. *Váhové koeficienty neurónovej siete sú náhodne vygenerované z intervalu [-1,1].*

Krok 2. *Polož* t:=1.

Krok 3. *S 50% pravdepodobnosťou deklaruj agenta ako prvého* X*-hráča a model ako druhého* O*-hráča (v opačnom prípade je agent deklarovaný ako druhý* O*-hráč a model ako prvý* X*-hráč). Na záver hry pomocou metódy TD(λ) opraví váhové koeficienty kognitívneho orgánu agenta.*

Krok 4. *Polož* t:=t+1.

Krok 5. *Ak* $t<t_{max}$*, potom pokračuj krokom 3, v opačnom prípade prejdi na krok 6.*

Krok 6. *Koniec algoritmu.*

Tento algoritmus bol použitý na naše simulačné výpočty, pričom konštanta λ=0.3, krok učenia α=0.1, a počet skrytých neurónov N_H=50. Získané numerické výsledky sú sumarizované na obr. 5.11, kde diagram znázorňuje frekvenciu výskytu víťazstva, prehry a remízy, ktoré v priebehu algoritmu boli počítané každých 100 epoch. Z diagramu vyplýva, že asi po 100000 epochách procesu učenia neurónovej siete metódou odmeny a trestu,

kognitívny aparát agenta už bol tak adaptovaný, že skoro 100 hier remizuje s použitým modelom

Záver

Uvedený subsymbolický prístup k riešeniu úloh, pre ktoré je len veľmi obtiažne zostrojiť ich efektívny model, poskytuje povzbudzujúce výsledky:

- V priebehu evolúcie agentov dochádza k emergencii stratégie hry.
- Pre hru backgamon získal Tesauro neurónovú sieť, ktorá je schopná hrať túto hru na veľmajstrovskej úrovni.
- Získané výsledky sú vynikajúcou ilustráciou subsymbolického prístupu k riešeniu zložitých úloh, ktoré sú ťažko riešiteľné technikami klasickej (symbolickej) umelej inteligencie.

V rámci subsymbolického prístupu, založenom na neurónových sieťach s dopredným šírením signálu a "učenia s odmenou a trestom" (reinforcement learning), je možné riešiť rôzne úlohy z robotiky, riadenia zložitých systémov, komplexných strategických hier, atď. bez nutnosti poznať ich model alebo databázu ich známych realizácií.

Literatúra

[1] Crowley, K., Siegler, R. S.: Flexible Strategy Use in Young Children's Tic-Tac-Toe". *Cognitive Science* **17** (1993), 531–561.

[2] Kvasnička, V.: Emergencia stratégie hry. AT&P Journal **12** (1999), 44-46.

[3] Lacko, P., Kvasnička, V.: Štúdium doskových hier pomocou neurónových sietí. In *Umelá inteligencia a konitívna veda III* (Kvasnička, V., PospíchaL, J., Návrat, P., Lacko, P., Varga, Ľ. (editori). STU Press, Bratislava, 2011, pp. 117-162.

[4] Návrat, P. (et al.): *Umelá inteligencia*. STU Press, Bratislava, 2002.

[5] Newell, A., Simon, H.A.: *Human problem solving*. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[6] Olazaran, M.: A Sociological History of the Neural Network Controversy. In Yovits, M. C. (ed.): *Advances in Coputers*, Vol. 37. Academic Press, Boston, MA, 1993.

[7] Russell, S. J., Norvig, P.: *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey, 2003.

[8] Sutton, R. S., Barto, A. G.: *Reinforcement Learning: An Introduction,* MIT Press, Cambridge, MA, 1998.

[9] Sutton, R. S.: Learning to Predict by the Methods of Temporal Differences. *Machine Learning* **3** (1988), 9-44.

[10] Tesauro, G.; Sejnowski, T. J.: A Parallel Network That Learns to Play Backgammon, Artificial Intelligence **39**(1989), 357-390

[11] Tesauro, G.: Practical Issues in Temporal Difference Learning. In Moody. J. E., Hanson, S. J., Lippmann, R. (Eds.): *Advances in Neural Information Processing Systems* 4, Morgan Kaufmann, San Fransisco, CA, 1992, pp. 259-266.

[12] Thorndike, E. L.: *The Fundamentals of Learning*. AMS Press Inc., New York, 1932.