## Cvičenia

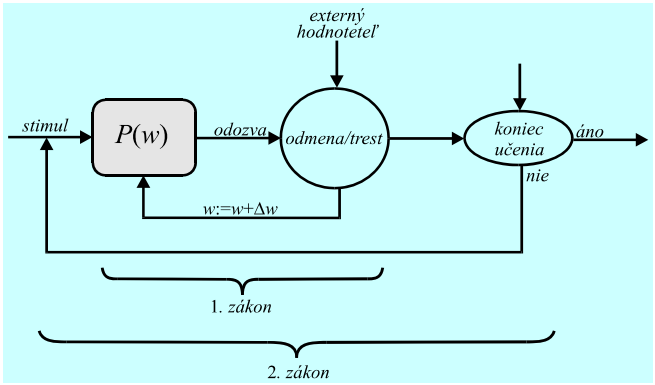
**Cvičenie 5.1.** Ako je definované podľa amerického psychológa Thorndike učenie s odmenou a trestom?

**Riešenie.** Thorndike (1887-1949) v 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.*

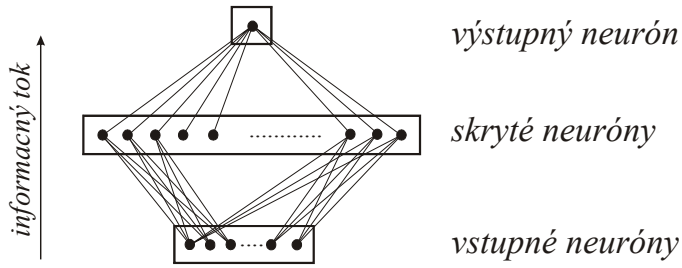
(2) **Zákon opakovaného používania**:*Požadované správanie je výsledkom častého používania dvojica stimul a odozva*

Potom metóda ***učenia s odmenou a trestom*** (reinforcement learning - temporal differences RL-TD(λ)) je učenie, ktoré je založené na opakovanom použití týchto dvoch zákonoch, ktoré tvoria teoretický základ behavioristického prístupu k učeniu.




**Cvičenie 5.2**. Ako sa používajú neurónové siete ako prostriedok pre rozhodovanie v hre?

**Riešenie. Nech a**gent s kognitívnym orgánom realizovaným pomocou doprednej troj-vrstvovej neurónovej siete

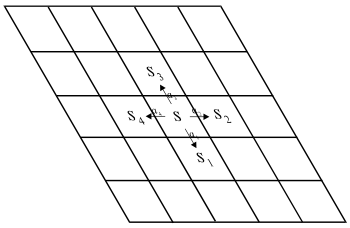



***Plasticita*** neurónovej siete sa realizuje pomocou učenia siete, t. j. systematickej zmeny váhových koeficientov tak, aby na výstupe neurónovej siete bola požadovaná aktivita ako odozva na daný vstup.

Študujme agentov, ktorý sú určený pomocou týchto dvoch množín: (1) Množina agentových **stavov** $\mathcal{S}=\{s_1, s_2, \ldots\}$ a (2) množina agentových **akcií**

$$\mathcal{A}=\{a_1, a_2, \dots\}$$

Akcie agentov sú interpretované ako funkcie, ktoré zobrazujú stavy agentov na seba, $s' = a(s)$.



Agent pomocou svojho ***kognitívneho orgánu*** (***prediktoru***) ohodnocuje svoje stavy reálným číslom (**predikciou**)

$$P(w): \mathcal{S} \to R$$

alebo explicitne

$$z = P(s; w)$$

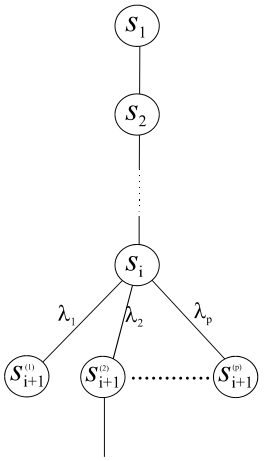
Zobrazenie $P$ – prediktor je parametrická funkcia. Hovoríme, že vykazuje **plasticitu** vzhľadom k parametrom $w$. Predpokladáme, že výber určitej akcie $a \in \mathcal{A}$, ktorá je aplikovaná na stav $s \in \mathcal{S}$ je riadený kognitívnym orgánom pomocou nasledujúceho postupu (učenia): Majme sekvenciu stavov agenta a jej ohodnotenie (pochvala - trest)

$$s_1, s_2, \dots, s_m, z$$

kde $z$ je ohodnotenie, ktoré vyjadruje skutočnosť, či sekvencia má alebo nemá požadovanú vlastnosť

$$z = \begin{cases} 1\,(sequence\ has\ the\ required\ \ property) \\ 0\,(otherwise) \end{cases}$$

Sekvencia stavov $s_1, s_2, \dots, s_m, z$ je zostrojená **kvázináhodne**, t.j. podsekvencia $s_1, s_2, \dots, s_m, z$ je rozšírená o ďalší stav $s_{i+1}$ na základe predikcie $z = P(s; w)$.



Kde posledný stav je vybraný pomocou maximálnej odozvy neurónovej siete

$$s_{i+1} = arg\ \max_j P\left(s_i, s_{i+1}^{(j)}; w\right)$$

Potom stratégia učenia neurónovej siete má cieľ adaptovať kognitívny orgán $P(w)$ tak, že všetky predikcie

$$P_1 = P(s_1; w), P_2 = P(s_2; w), ..., P_m = P(s_m; w)$$

sú rovnaká, ako externé ohodnotenie celej sekvencie číslom $z = P_{m+1}$. **Poznámka**: Používame takú stratégiu učenia, že ak je výsledná sekvencia vyhodnotená ako úspešná (neúspešná), potom všetky stavy z tejto sekvencie sú tiež úspešné (neúspešné).

**Cvičenie 5.3.** Aký je klasický prístup k učeniu neurónovej siete, aby vedela vyhodnocovať prechod z jedného stavu do druhého stavu pri konštrukcii stromu riešenia?

**Riešenie**. Klasický prístup k adaptácii neurónových sietí je založený na použití tzv. tréningovej množiny, ktorá obsahuje dvojice vstupu (špecifikujúceho daný stav – pozíciu) a požadovaného výstupu (hodnotenie pozície číslom z intervalu (0,1))

$$\mathcal{A}_{train} = \{ \boldsymbol{x} = (x_1, x_2, ..., x_n) / z_{req} \}$$

Použitá neurónová sieť v tomto klasickom prístupe je parametrické zobrazenie $n$-rozmerných vektorov na reálne číslo z otvoreného intervalu (0,1)

$$G(w): R^n \to (0,1)$$

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

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

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

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

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

kde $\alpha$ je kladný parameter nazývaný „krok učenia“. Výpočet parciálnych derivácií výstupných aktivít vzhľadom k parametrom neurónovej siete je vykonaný rekurentnou metódou „backpropagation error“. Tento „klasický“ prístup k adaptácii neurónovej siete, ktorá sa používa na hodnotenie pozícií hier je veľmi ťažkopádny . 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 $z_{\text{req}}$. Pre hru TTT, 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.

Niekoľko záverečných poznámok k tomuto priamočiaremu použitiu neurónovej siete k implemenmtácii algoritmu pre symetrickú hru dvoch hráčov. 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.

**Cvičenie 5.4.** Popíšte adaptáciu neurónovej siete pomocou metódy odmeny a trestu vo verzii „temporal difference TD(λ)“.

**Riešenie.** Základné princípy „reinforcement learning“ 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, 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.

Predpokladajme, že poznáme sekvenciu pozícií a jej ohodnotenie reálnym číslom, ktoré odpovedajú pozíciám daného agenta – hráča, ktoré musel ohodnocovať číslom $z$

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

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

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

Vytvoríme účelovú funkciu

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

Cieľom učenia sú 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 $z_{\text{reward}}$. Rekurnetná formula pre obnovu váhových koeficientov má tvar

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

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

kde $z_t=G(P_t,w)$ je ohodnotenie $t$-tej pozícii $P_t$ pomocou neurónovej sieti - kognitívneho orgánu číslom $z_t$. Našim cieľom bude, aby všetky pozície zo sekvencie boli ohodnotené rovnakým číslom $z_{\text{reward}}$, ktoré nám špecifikuje, či hra daného hráča pozostávajúca zo sekvencií bola víťazná, remízová, alebo viedla k prehre. Výraz v zátvorke prepíšeme do tvaru

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

kde sme použili algebraickú identitu

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

Potom dostaneme

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

$$= \alpha \sum_{t=1}^{m}\left(z_{t+1}-z_t\right)\sum_{k=1}^{t}\frac{\partial z_t}{\partial w}$$

Inkrement obnovovacej formule váhového koeficienta má tvar

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

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

Tento dôležitý výsledok môže byť "zovšeobecnený" na formulu, ktorá tvorí základ TD(λ) rodiny učiacich metód “s odmenou a trestom”

$$w := w + \sum_{t=1}^{m} \Delta w_t$$

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

$$\Delta w_t = \alpha\left(P_{t+1}-P_t\right)e_t\left(\lambda\right)$$

$$e_{t+1}\left(\lambda\right) = \lambda e_t\left(\lambda\right) + grad_w P_{t+1}$$

$$e_1\left(\lambda\right) = grad_w P_1$$

**Cvičenie 5.5.** (Námet pre esej) Pre dané jednoduché bludisko obsahujúce 10-20 uzlov, ktoré sú označené symbolmi zostrojte pomocou učenia s odmenou a trestom takú doprednú neurónovú sieť, ktorá je schopná efektívne riadiť agenta pri pohybe bludiskom.

**Cvičenie 5.6.** (Námet pre zložitejšiu esej) Emergenciu stratégiu hry piškvorky študujte pre populáciu agentov s kognitívnym orgánom (implementovaným pomocou doprednej neurónovej siete) tak, že v populácii agentov prebieha neustály turnaj pre náhodne vybranú dvojicu agentov $A_1$ a $A_2$, po skončení danej hry agenti $A_1$ a $A_2$ si adaptujú pomocou TD(λ) metódy svoje neurónové siete, pričom v populácii víťaz nahradí v populácii porazeného agenta; ak agenti remizovali, potom oba agenti v populácii zostávajú. Nárast efektívnosti hry agentov sledujte tak, že v každom n-tom kroku (nech n = 100) posledný víťaz hrá 10 hier proti agentovi riadenom formálnymi pravidlami.