National Conference with International Participation
ENGINEERING MECHANICS 2001
Svratka, Czech Republic
May 14 - 17, 2001

REINFORCEMENT LEARNING MODEL: CONTROL OF NONLINEAR AND UNSTABLE PROCESSES

Tomáš Březina[1], Zdeněk Ehrenberger[2], Ctirad Kratochvíl[3]

Summary: *Some experiences with using of the reinforcement learning model at control of nonlinear unstable processes are published in this paper. Control process is characterized by extensive depth in such cases, so the learning is computationally very demanding. We propose both using of nonlinear grid of the Q-function approximation table and also using of the learning conception "by expert observation". Learning off (optimal) control policy is not based on blind searching a state space, but it is in progress with the help of further component, that is able to control the process. Problems are studied on active magnetic bearing one-mass model.*

1. ÚVOD

Jedním z cílů umělé inteligence je získat nástroje k tvorbě "inteligentních agentů" samostatně dosahujících cílů ve složitých prostředích. Základní výhodou učení je schopnost zlepšovat chování agentů v čase.

Bez znalosti požadovaného chování agenta může být velmi obtížné získat i malou množinu příkladů (tréninkových faktů) tohoto chování. Disponovat množinou tréninkových faktů je ale nutné, je-li problém formulován jako úloha učení s učitelem.

Problematiku však lze přirozeně formulovat jako optimalizační úlohu, kde je optimální chování definováno jako chování, které průběžně extremalizuje nějakou cílovou funkci agentova chování, jejíž odhad je průběžně zpřesňován. Zpřesňování odhadu probíhá na základě jednoduchého ohodnocování okamžitého výkonu agenta. Takové úlohy se nazývají úlohy opakovaně posilovaného učení (reinforcement learning, RL). Konvenční RL architektury jsou ale většinou dost pomalé na to, aby byly pro řešení mnoha úloh prakticky použitelné.

V příspěvku je hodnocen vliv volby rastru na rychlost učení RL architektury postavené na tabulce a dále vliv konceptu učení "pozorováním experta". Naučení (optimální) řídící strategie zde není založeno na pouze slepém prozkoumávání stavového prostoru prostředí, ale probíhá za pomoci dalšího členu schopného proces řídit.

[1] Tomáš Březina, Ing.RNDr.CSc.,VUT v Brně, FSI ÚAI, Technická 2, 616 69 Brno, brezina@uai.fme.vutbr.cz
[2] Zdeněk Ehrenberger, Prof.Ing.DrSc., VUT v Brně, FSI ÚMT CM, Technická 2, 616 69 Brno, ehrenberger@uvssr.fme.vutbr.cz
[3] Ctirad Kratochvíl, Prof.Ing.DrSc., VUT v Brně, FSI ÚAI, Technická 2, 616 69 Brno, kratochvil@umt.fme.vutbr.cz

Problematika je studována na simulaci řízení jednohmotového modelu aktivního magnetického ložiska (AML).

2. POUŽITÁ KONCEPCE

Q-učení je pravděpodobně nejpopulárnější a nejefektivnější RL algoritmus bez modelu. Optimální Q-funkce je definována jako $Q^*(s_i,u)=\max_\pi Q_\pi(s_i,u), \forall s_i\in S, \forall u\in U$, kde $Q_\pi(s_i,u)=r(s_i,u)+\sum_{k=i+1}^{\infty}\gamma^{k-i}r(s_k,\pi(s_k))$, $0<\gamma<1$, [2]. Optimální strategie je potom $\pi^*(s)=\arg\max_u Q^*(s,u)$ (obdobně $\pi(s)=\arg\max_u Q(s,u)$). Snad nejdůležitější vlastností Q-učení je to, že Q-hodnoty konvergují ke Q^* nezávisle na chování agenta. Bohužel však konvergence může být velmi pomalá.

Provedl-li agent v čase t_k akci u_k, která převedla stav prostředí z s_k do s_{k+1} (v čase t_{k+1}) je jednou z možností realizace kroku učení

$$\Delta e(s,u)=\begin{cases}(\gamma\lambda-1)e(s,u)+1, & \text{pro } (s=s_k)\wedge(u=u_k)\\ (\gamma\lambda-1)e(s,u), & \text{jinak}\end{cases},\ \forall s\in S,\forall u\in U,$$
$$\Delta Q(s,\,u)=\alpha\Big(r(s_k,\,u_k)+\gamma\max_{\mathbf{u}_{k+1}}Q(s_{k+1},\,u_{k+1})-Q(s_k,\,u_k)\Big)e(s,u),\ \forall s\in S,\forall u\in U. \tag{1}$$

Učení "pozorováním experta" vychází z toho, že je pro daný stav známa přiměřená akce převzatá od experta. Mělo-li by být $\pi(s)=u'$, potom jednoduchou volbou

$$Q(s,\,u')=\max_{u\neq u'}Q(s,\,u)+\varepsilon,\ \varepsilon>0 \tag{2}$$

bude zajištěn výběr akce s'.

Jako modelu prostředí je použito jednoduchého jednohmotového modelu AML [1]

$$\begin{aligned}m\ddot{x}(t)+b\dot{x}(t)+F_m\big(x(t),I(t)\big)&=F_e(t)\\ RI(t)+L\dot{I}(t)&=u(t)\end{aligned},\ x(t)=x_0,I(t)=I_0,\dot{x}(t)=\dot{x}_0 \text{ pro } t\le 0, \tag{3}$$

kde m je hmotnost hmotného bodu, $x(t)$ výchylka hmotného bodu, $i(t)$ značí proud procházející ložiskovými elektromagnety, R a L jsou odpor a indukčnost cívek elektromagnetů, síla $F_e(t)$ zavádí rušení způsobené nevyvážeností a zatížením rotoru, $u(t)$ představuje napětí na elektromagnetech, b viskózní tlumení (zavádí vliv okolního prostředí na model), a konečně $F_m(x(t),i(t))$ je síla od magnetického ložiska působící na hmotný bod.

Úkolem agenta je, aby svým chováním (generováním akcí) dostal výchylku hmotného bodu do intervalu $\langle -x_r, x_r\rangle$ a udržel ji v něm. Agent tak představuje řídící člen AML.

3. IMPLEMENTACE Q-FUNKCE

Stav prostředí s je chápán jako aritmetický vektor jistých veličin prostředí $\mathbf{s}=(s^1,\ldots,s^n)$, Podobně akce u jako aritmetický vektor řídících veličin $\mathbf{u}=(u^1,\ldots,u^m)$. Nechť $\{g^i_j\}_{j=0}^{p_i}$, je posloupnost uzlových bodů rastru i-té veličiny prostředí (konstituující stav) z s^i taková, že $g^i_j<g^i_{j'}\Leftrightarrow j<j'$, $j,\,j'=0,\ldots,p_i$. Hodnotou indexu reprezentujícím

hodnotu i-té veličiny prostředí vzhledem k rastru nazveme takový index j, pro který $g_j^i \le s^i < g_{j+1}^i$. Zcela analogicky je zavedena reprezentace akce skládající se z řídících veličin s uzlovými body rastru $\left\{h_j^i\right\}_{j=0}^{r_i}$. Funkce $Q(\mathbf{s},\mathbf{u})$ je potom reprezentována tabulkou, jako

$$Q(\mathbf{s},\mathbf{u}) = q_{j^1 j^2 \ldots j^{n+m}},$$

kde j^i reprezentuje hodnotu i-té veličiny prostředí (resp. $i-n$-té řídící veličiny).

Přiměřenost aproximace funkce $Q(\mathbf{s},\mathbf{u})$ tabulkou bude zřejmě ovlivněna konkrétní volbou uzlových bodů rastru. Bude-li polí rastru málo, půjde zřejmě o příliš hrubou aproximaci (která nepovede ani na suboptimální řídící strategii), bude-li rastr příliš jemný, stane se neúnosným požadavek modifikací Q-hodnot nad značně vysokým počtem polí. V prvém případě je vyloučeno dosažení funkce $Q^*(\mathbf{s},\mathbf{u})$ s dostatečnou přesností, v druhém případě toto není dosažitelné bez zdlouhavého výpočtu.

Dalším faktorem je to, že "kvalita" strategie řízení bude zřejmě záviset i na velikosti a poloze (tj. uzlových bodech) jednotlivých polí rastru. Pro množinu stavů s požadavkem "jemnějších" řídících zásahů bude třeba jemnějšího rastru, než v oblastech bez této podmínky. Z hlediska dosažení strategie řízení s jistou hladinou úspěšnosti během určitého počtu průchodů výpočetní procedurou tedy lze očekávat existenci optimálního rastru, jak co do počtu polí, tak co do jejich velikostí.

4. PROVEDENÉ EXPERIMENTY

Jsou popsány dvě sady experimentů: s lineárním rastrem ($g^i_{j^i+1} - g^i_{j^i} = konst$, $j_i = 0,\ldots,p_i$, $i = 1,\ldots,n$ a opět analogicky pro $\left\{h_j^i\right\}_{j=0}^{r_i}$) a s nelineárním rastrem, kdy jsou podmínky lineárního rastru aspoň v jednom případě porušeny.

Při všech prováděných experimentech byly akce agenta byly brány z množiny $\{-U_{\max}, 0, U_{\max}\}$. Počáteční podmínky x_0, $\dot{x}_0$, I_0 byly voleny náhodně, ale pouze takové, pro které je soustava "řiditelná", tj. použití napětí $U_{\max}$ (s vhodným znaménkem) zaručuje existenci takového času τ, pro který $x(\tau) = 0$.

Posilovací funkce agenta r byla definována jako

$$r = \begin{cases} 1, & \text{pro } (|x_k| \le x_r) \wedge (|\dot{x}_k| \le \dot{x}_r) \\ 0, & \text{pro } ((x_r < |x_k| < x_{\max}) \vee (\dot{x}_r < |\dot{x}_k|)) \wedge (|x_k| > |x_{k+1}|) \\ -1, & \text{jinak} \end{cases},$$

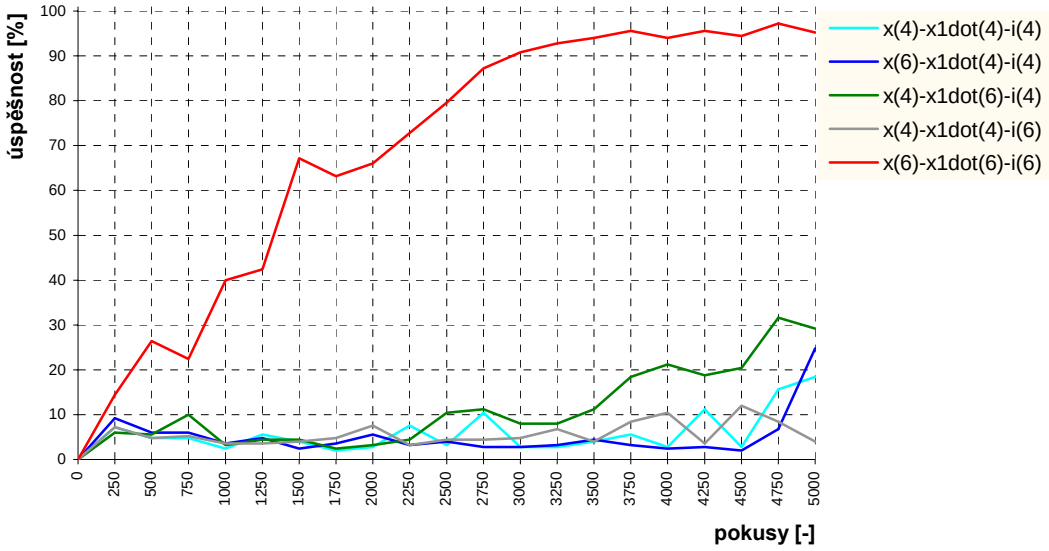
kde x_r a $\dot{x}_r$ jsou velikosti okamžité výchylky a rychlosti uvnitř kterých je hmotný bod pokládán za stabilizovaný.

Učení bylo chápáno jako úspěšné, jestliže se během něj soustava řídícími zásahy dostala do stavu, ve kterém platí $(|x_k| \le x_r) \wedge (|\dot{x}_k| \le \dot{x}_r)$. Učení bylo pokládáno za neúspěšné, jestliže se soustava dostala do stavu, kdy $|x_k| = x_{\max}$. Učení probíhalo tak dlouho, dokud nebyl zaznamenán buď úspěch, nebo neúspěch. Úspěšnost znamená počet úspěšných učení (pokusů) vztažený na počet všech provedených učení (pokusů).

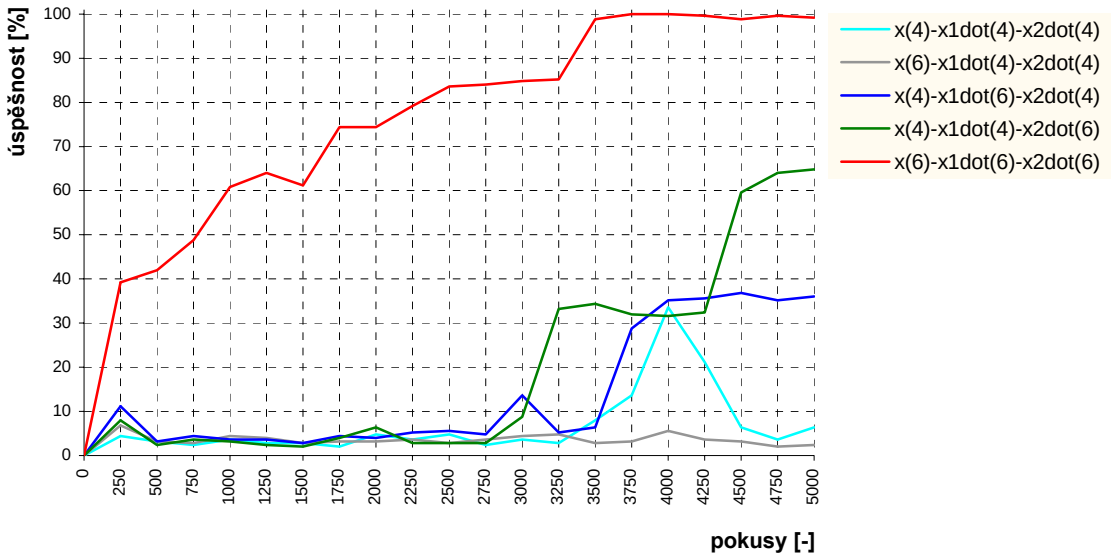
Pro experimenty bylo použito hodnot parametrů modelu podle [1] ($m = 5,65\,[\mathrm{kg}]$, $R = 2\,[\Omega]$, $L = 5 \cdot 10^{-3}\,[\mathrm{H}]$, $b = 20\,[\mathrm{kgs^{-1}}]$, $c_1 = 1289 \cdot 10^3\,[\mathrm{kgs^{-2}}]$, $c_2 = 454\,[\mathrm{kgms^{-2}A^{-1}}]$, $c_3 = 3184 \cdot 10^9\,[\mathrm{kgm^{-2}s^{-2}}]$, $c_4 = 841 \cdot 10^6\,[\mathrm{kgm^{-1}A^{-1}s^{-2}}]$, $c_5 = 38 \cdot 10^3\,[\mathrm{kgs^{-2}A^{-2}}]$, $c_6 = 30\,[\mathrm{kgms^{-2}A^{-3}}]$). Hodnoty x_r a $\dot{x}_r$ byly zvoleny $x_r = 10.10^{-6}\,[\mathrm{m}]$, $\dot{x}_r = 20.10^{-3}\,[\mathrm{ms^{-1}}]$, $U_{\max} = 80\,[\mathrm{V}]$.

Počáteční testy měly za úkol zjistit, jak hrubý rastr stavových veličin je pro učící algoritmus kritický a dále, jeví-li se jako vhodnější z hlediska úspěšnosti učení stav prostředí definovaný jako $(x_k, \dot{x}_k, I_k)$ nebo jako $(x_k, \dot{x}_k, \ddot{x}_k)$. Z výsledků experimentů vyplynulo, že pro obě uvažované definice stavů prostředí je hranice schopnosti učení určena rozdělením každé ze stavových proměnných na 6 intervalů ($n = 3$, $p_1 = p_2 = p_3 = 6$) a dále, že se pro další experimenty jeví jako nadějnější definice $(x_k, \dot{x}_k, \ddot{x}_k)$, viz. grafy na obr. 1 a obr. 2.

V legendách následujících grafů je použito označení x pro x, x1dot pro $\dot{x}$, x2dot pro $\ddot{x}$ a i pro I. Připojené číslo značí počet dílků rastru dané veličiny.



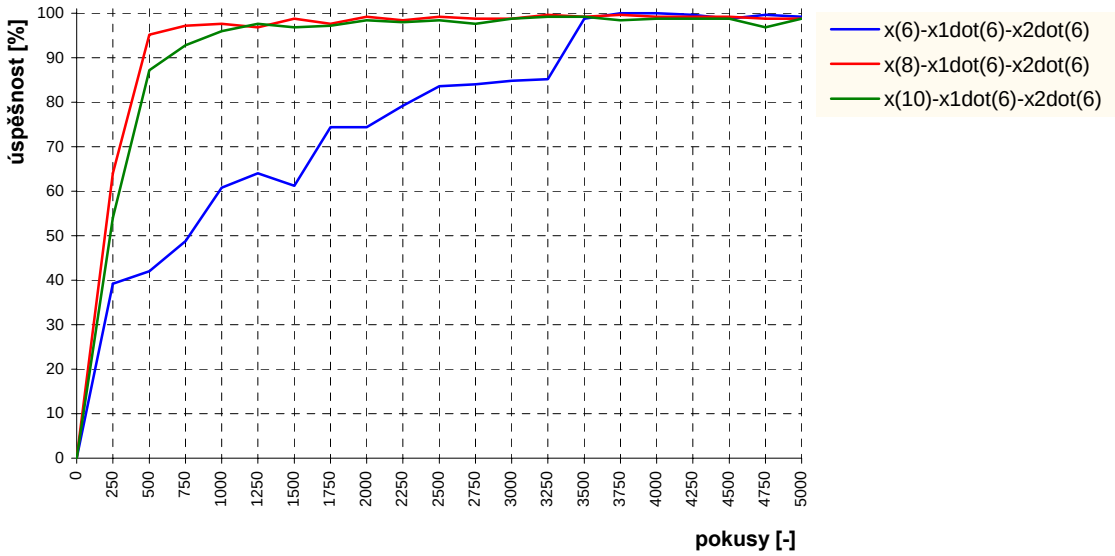
Obr.1: Minimální rastr, stavy prostředí $(x_k, \dot{x}_k, I_k)$



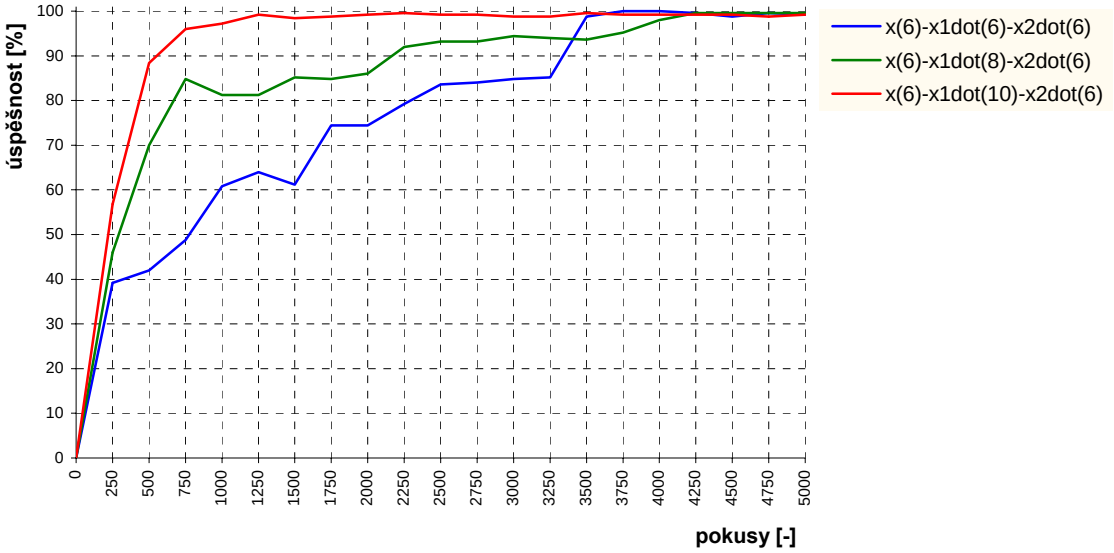
Obr.2: Minimální rastr, stavy prostředí $(x_k, \dot{x}_k, \ddot{x}_k)$

4.1. Lineární rastr

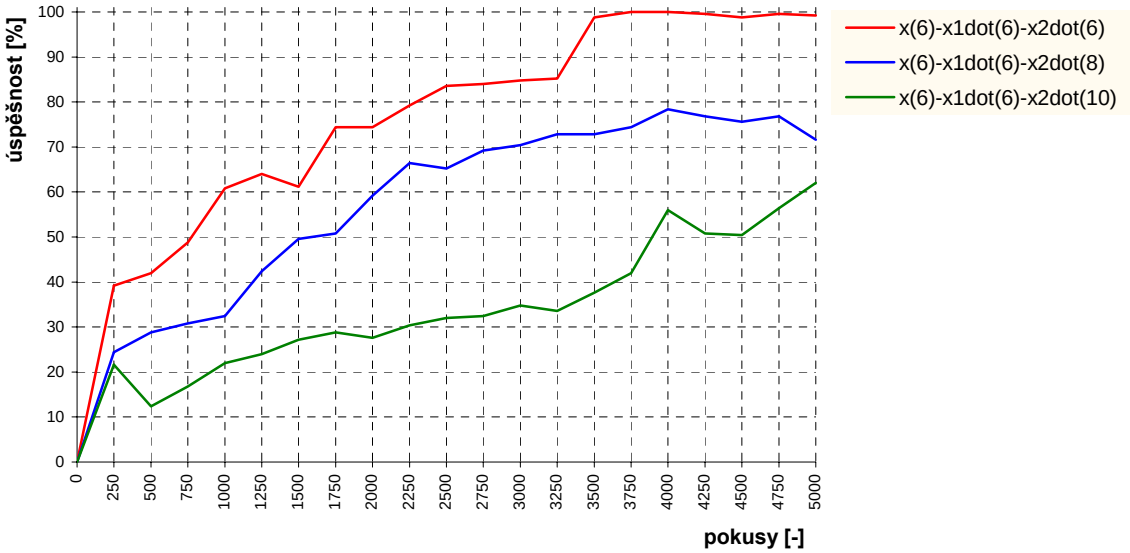
Experimenty prováděné v dalších etapě se týkaly optimalizace počtu intervalů rastru jednotlivých stavových veličin. Jako stavový prostor byl, na základě výsledků předchozích experimentů, brán případ $(x_k, \dot{x}_k, \ddot{x}_k)$. Výsledky jsou shrnuty na obr. 3-5.



Obr.3: Vliv volby lineárního rastru výchylky na průběh učení



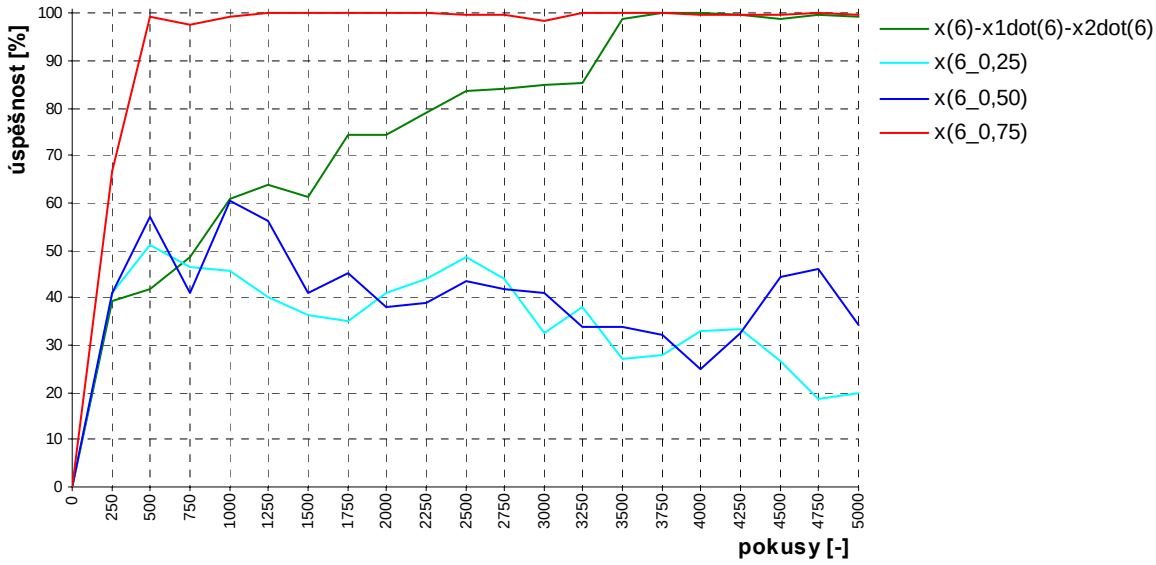
Obr.4: Vliv volby lineárního rastru rychlosti na průběh učení



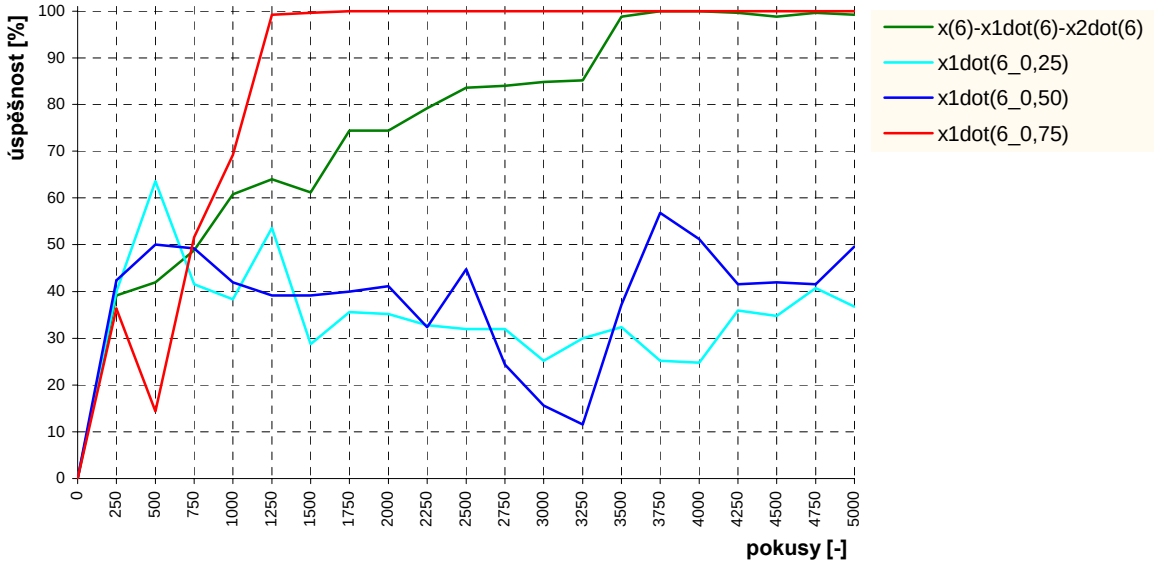
Obr.5: Vliv volby lineárního rastru zrychlení na průběh učení

4.2. Nelineární rastr

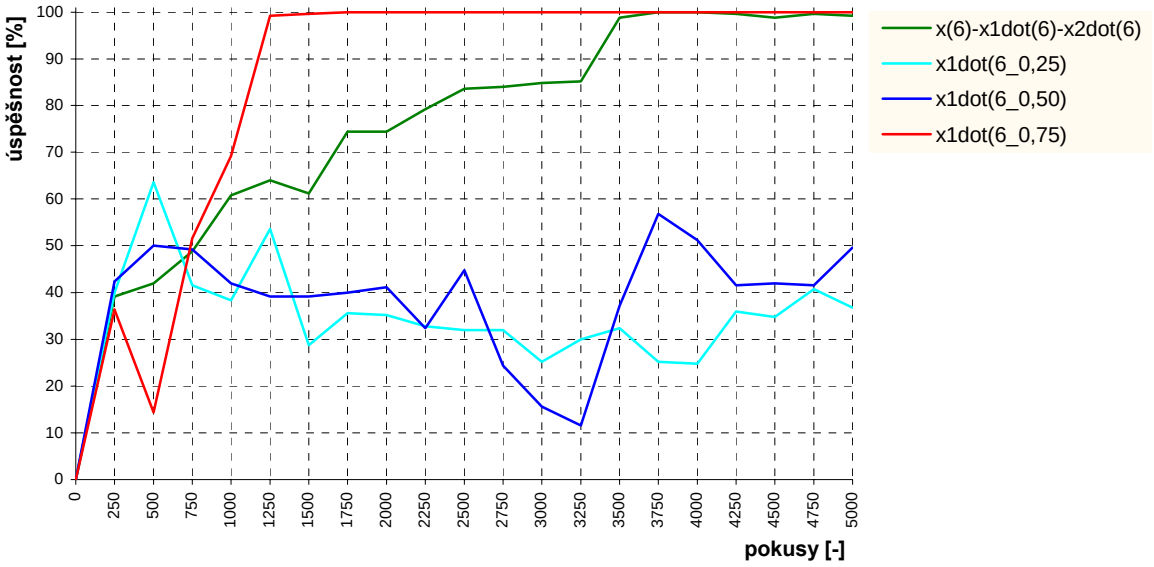
Další část experimentů se týkala nelineárního dělení rastrů, které by při zachování počtu dílků rastrů jednotlivých veličin mohlo přinést lepší průběh učení. Nelinearita rastru byla nastavována hodnotou parametru $a \in \langle 0,1 \rangle$ úměrně členu $a^{|\Delta i|}$, kde $|\Delta i|$ značí vzdálenost indexu *i*–tého dílku rastru od indexu středového dílku (tj. $a = 1$ odpovídá lineárnímu rastru). V legendách následujících grafů je původní značení rozšířeno o hodnotu tohoto parametru. Výsledky jsou shrnuty na obr. 6-8.



Obr.6: Vliv volby nelineárního rastru výchylky na průběh učení



Obr.7: Vliv volby nelineárního rastru rychlosti na průběh učení



Obr.8: Vliv volby nelineárního rastru zrychlení na průběh učení

4.3. Učení "pozorováním experta"

Experimenty s učením "pozorováním experta" nejsou dosud uzavřeny. Dosavadní výsledky naznačují, že "pozorování experta" podstatně zvyšuje úspěšnost v počátečních fázích učení (při cca 150 pokusech až na 70%), ovšem při přechodu na standardní Q-učení může dojít, i při zvyšování počtu pokusů, ke stagnaci úspěšnosti na hladině asi 95% nebo dokonce k jejímu zhoršení.

5. ZÁVĚR

Metoda využívá schopnost učení postupně zlepšovat chování AML. Provedené simulace prokázaly existenci (sub)optimálního rastru použitého v aproximaci Q-funkce. Při jeho použití dochází k výraznému zrychlení procesu učení. Prostředí reprezentované modelem AML je řiditelné tabulkou (s prohledávacím mechanismem) obsahující $6\times6\times6\times3$ hodnoty (a funkcí max). Je pravděpodobné, že současné zavedení nelineárního rastru pro všechny veličiny prostředí povede k dalšímu zlepšení procesu učení.

Rovněž učení "pozorováním experta" může přinést podstatné zrychlení procesu učení, při porovnání se standardním učením (používajícím slepého prohledávání stavového prostoru). Výrazně zvyšuje počáteční úspěšnost učení, ale dlouhodobě může vést ke stagnaci nebo dokonce zhoršení úspěšnosti. Toto konstatování však vychází z dosud neuzavřených experimentů.

6. PODĚKOVÁNÍ

Tato práce je podporována projekty GAČR č. 101/00/1471 a CEZ J22/98:261100009. Dík patří Zdeňku Pavlusovi za provedení většiny zmiňovaných experimentů.

7. LITERATURA

1. Půst L.: Dynamic stability of magnetic bearing, Engineering Mechanics 97, Svratka, pp. 57-62, 1997
2. F Watkins C., Dayan P.: Q-learning, Machine Learning, 8(3), pp. 279-292, 1992