INŽENÝRSKÁ MECHANIKA 2005
NÁRODNÍ KONFERENCE
s mezinárodní účastí
Svratka, Česká republika, 9. - 12. května 2005

NONSTATIONARY SYSTEM CONTROL USING Q-LEARNING

S. Věchet*, J. Krejsa+

Summary: *Q-learning is the most popular and effective version of Reinforcement Learning algorithms. In this paper we discuss the possibility of control of a non-stationary system by Q-learning. The non-stationary system is represented by simple inverted pendulum simulation model with variable pendulum length.*

1. Úvod

Jedním z nejrozšířenějších algoritmů, na jehož základě je možno realizovat robustní řídicí člen, je v současné době Q-učení, které patří do skupiny algoritmů opakovaně posilovaného učení (Reinforcement Learning – RL). Opakovaně posilované učení je obecně založeno na vzájemném vztahu agenta a prostředí, kdy agent svými akcemi ovlivňuje prostředí a na základě vhodně ohodnocené změny stavu prostředí generuje akci vedoucí k požadovanému stavu prostředí. Hlavní výhodou použití Q-učení je to, že pro úspěšné naučení se regulovat danou soustavu není nutně vyžadován model soustavy a dále není nutné předem znát pro daný stav optimální akci. Q-učení je navíc oblíbené pro svou jednoduchou implementaci.

Ve své podstatě je Q-učení navrženo tak (jak bude ukázáno dále) aby konvergovalo k nalezení optimální strategie řízení, i když konvergence může být značně pomalá. Této vlastnosti lze s úspěchem využít na konstrukci řídicího členu pro nestacionární soustavu. Pokud soustava mění své vlastnosti v řádově delším časovém intervalu než je rychlost učení Q-učení, je možno zajistit aby se daný řídicí člen neustále přizpůsoboval řízené soustavě. V praxi je možno každou soustavu označit za nestacionární, neboť zde dochází k opotřebení a stárnutí jednotlivých mechanických částí a tím i ke změnám v parametrech soustavy. Tento článek ukazuje vlastnosti metody Q-učení ve spojitosti s řízením nestacionární soustavy, která je representována jednoduchým modelem inverzního kyvadla. Nestacionarita tohoto modelu je zajištěna změnou délky ramene kyvadla v průběhu učení.

2. Opakovaně posilované učení

Klasický model opakovaně posilovaného učení je tvořen agentem a prostředím. V každém časovém okamžiku t se prostředí nachází ve stavu s_t. Agent má k dispozici množinu akcí, kterými stav prostředí ovlivňuje. Poté, co agent provede akci a_t, způsobí změnu stavu prostředí na stav s_{t+1}. Jednou z možností jak specifikovat požadované chování agenta je definovat funkci okamžitého posílení $r(s_t,st_{+1},a_t)$, která určuje konkrétní odměnu/pokutu za přechod ze stavu s_t do stavu s_{t+1} při použití akce a_t. Dlohodobý cíl agenta je definován jako

* Ing. Stanislav Věchet, PhD., VUT Brno, Technická 2, 616 69, Brno, Czech Republic, email: vechet.s@fme.vutbr.cz

+ Ing. Jiří Krejsa, PhD., ÚT AV ČR, pobočka Brno, Technická 2, 616 69, Brno, Czech Republic; tel: +420 541142885, email: jkrejsa@umt.fme.vutbr.cz

funkce okamžitých odměn, například kumulativní srážková odměna (cumulative discount reward)

$$\sum_{t=0}^{\infty} \gamma^t r(s_t, s_{t+1}, a_t) \quad (1)$$

kde $0 \leq \gamma < 1$ je srážkový faktor řídicí relativní důležitost krátkodobých a dlouhodobých odměn. Strategii agenta (pravidlo pro výběr akce *a* v daném stavu *s*) lze formálně zapsat ve tvaru $a = \pi(s)$ a cílem opakovaně posilovaného učení je najít optimální strategii π^*, která maximalizuje kumulativní srážkovou odměnu. Pro účel nalezení optimální strategie zavádíme hodnotovou funkci $f^{\pi}(s)$ strategie π, která udává oěkávanou kumulativní srážkovou odměnu při počátečním stavu *s* a použití strategie π.

3. Q-učení

V Q-učení je hodnotová funkce $f^{\pi}(s)$ nahrazena funkcí akční hodnoty *Q(s,a)*. Hodnota této funkce udává očekávanou kumulativní srážkovou odměnu při provedení akce *a* ve stavu *s* a při a při následném pokračování v dané strategii. Konkrétní hodnotou hodnotové funkce je potom maximum Q-hodnot pro daný stav:

$$f(s) = \max_a Q(s,a) \quad (2)$$

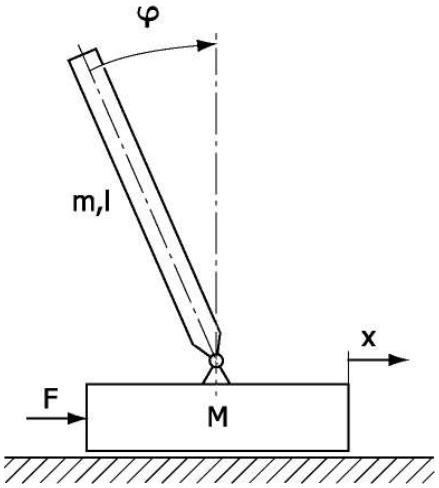
Q-funkce může být implementována různými zmpůsoby, v použitém případě implementací tabulkou je přepočtový vztah pro Q-funkci:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left[r(s_t, a_t, s_{t+1}) + \gamma \max_{a_t} Q(s_{t+1}, a_t) - Q(s_t, a_t) \right] \quad (3)$$

Pravděpodobně nejdůležitější vlastností Q-učení je to, že Q-hodnoty konvergují k optimální Q-funkci nezávisle na chování agenta (nezáleží na způsobu procházení kombinací jednotlivých stavů akcí). Q-hodnoty konvergují s pravděpodobností jedna v případě, že jsou v průběhu učení všechna uzlová místa navštívena nekonečněkrát (každá akce je v každém stavu vykonána nekonečně krát během nekonečného množství kroků), konvergence tedy může být značně pomalá.

4. Simulační model inverzního kyvadla

Simulační model inverzního kyvadla byl použit pro svou jednoduchost a proto, že jsou dostatečně známy jeho vlastnosti a je možné věnovat pozornost spíše studování vlastností metody Q-učení. Dále byl použit z toho důvodu, že ze své podstaty se jedná o model nestabilní a je nutné jej pro udržení dané polohy neustále řídit. Použitý simulační model inverzního kyvadla je schematicky znázorněn na obrázku 1.



Obr. 1 Simulační model inverzního kyvadla

a lze jej popsat rovnicemi:

$$\begin{aligned} &(M+m)\,\ddot{x} + ml\ddot{\varphi}\cos\varphi - ml\dot{\varphi}^2\sin\varphi = F \\ &\frac{12}{13l}\left(g\sin\varphi + \ddot{x}\cos\varphi\right) = \ddot{\varphi} \end{aligned} \qquad (4)$$

kde M je hmotnost vozíku, m hmotnost kyvadla, l délka kyvadla, F síla působící na vozík, g tíhové zrychlení, x souřadnice vozíku, φ úhel odklonu kyvadla od vertikální osy. Při simulacích byly zanedbány pasivní odpory.

5. Simulační přístupy

Experimenty byly prováděny následujícím způsobem: nejprve bylo vygenerováno n počátečních stavů, tyto stavy byly použity během testování a bylo sledováno, zda po stanovený počet kroků dokáže regulační člen udržet výchylku kyvadla v určeném intervalu. Pokud toho bylo dosaženo, byl pokus vyhodnocen jako úspěšný.

V experimentech bylo použito následujících parametrů modelu inverzního kyvadla: *M=0.2kg, m=0.1kg, g=9.81kgms*$^{-2}$*, l=0.1-2m.* Délka kyvadla byla měněna v daném intervalu během učení a byl sledován vliv velikosti změny na kvalitu učení.

Simulační experimenty byly hodnoceny následujícím způsobem. Pro hodnocení výsledků simulací byl zaveden termín *pokus*. Provedeni jednoho pokusu představuje simulaci procesu řízení, která trvá tak dlouho, dokud není splněna jedna z následujících podmínek:

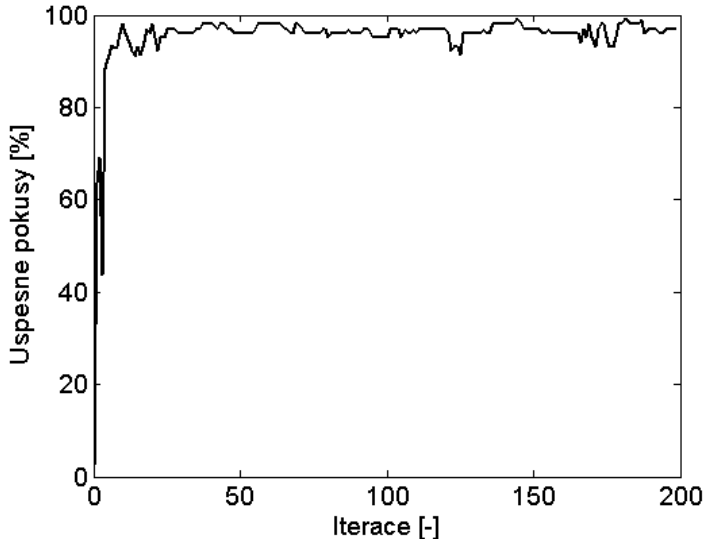
- Soustava dosáhne nekorektního stavu, tedy některý z parametrů soustavy je mimo vymezený rozsah. V takovém případě hovoříme o *neúspěšném pokusu*.
- Není vyčerpán stanovený počet řidicích rozhodnutí (akčních zásahů), nazvaný *maximální délka pokusu*. V tomto případě hovoříme o *úspěšném pokusu*.

Délkou pokusu je označen počet řídicích rozhodnutí provedených během pokusu. Délka pokusu odpovídá počtu úspěšných řídicích rozhodnutí. *Úspěšnost pokusu* je délka pokusu vztažená k maximální délce pokusu. Průměrná délka pokusu a procento úspěšných pokusů byly stanovovány vždy ze 100 pokusů, které se lišily pouze v počátečních stavech daného modelu inverzního kyvadla.

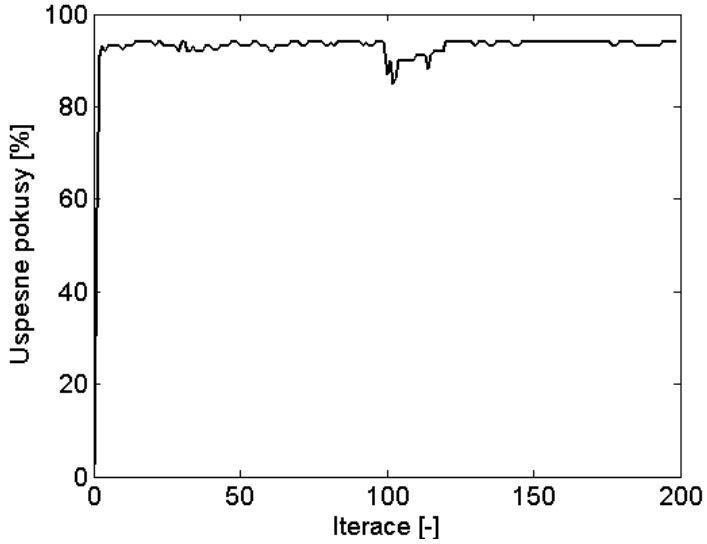
Počáteční stavy byly voleny z takových stavů soustavy, pro které daný model neuskutečnil během daného časového intervalu použitím libovolné akce z množiny akcí přechod do nekorektního stavu soustavy. Takto byly hrubě odhadnuty řiditelné stavy soustavy.

6. Praktické experimenty

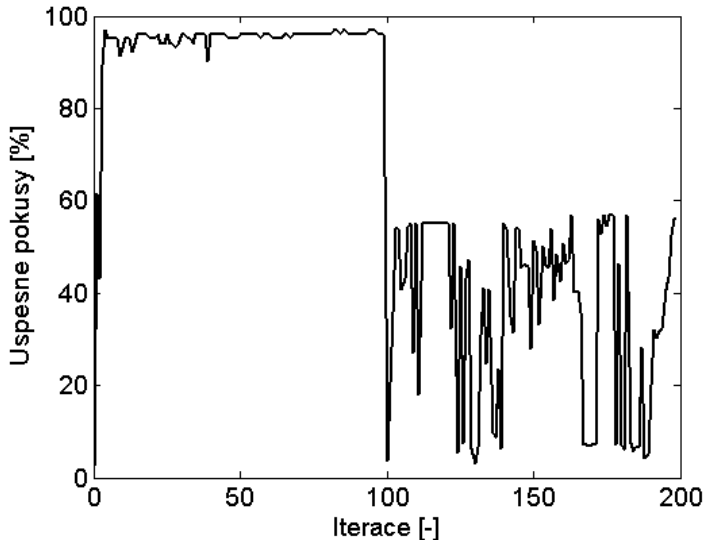
V této části jsou ukázány praktické výsledky z provedených experimentů. Obrázky 2 až 5 ukazují vliv velikosti změny délky ramene kyvadla na procento úspěšných pokusů. Délka kyvadla byla měněna skokově v průběhu učení. Na těchto grafech je ukázán vliv skokové změny délky kyvadla vždy ve 100. iteraci.



Obr. 2 Vliv zvětšení délky kyvadla z 1 na 2m



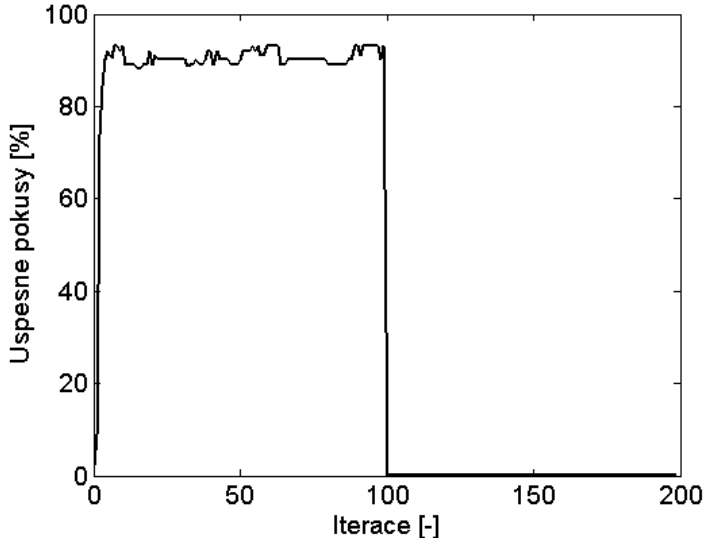
Obr. 3 Vliv zmenšení délky kyvadla z 1 na 0,6m



Obr. 4 Vliv zmenšení délky kyvadla z 1 na 0,2m

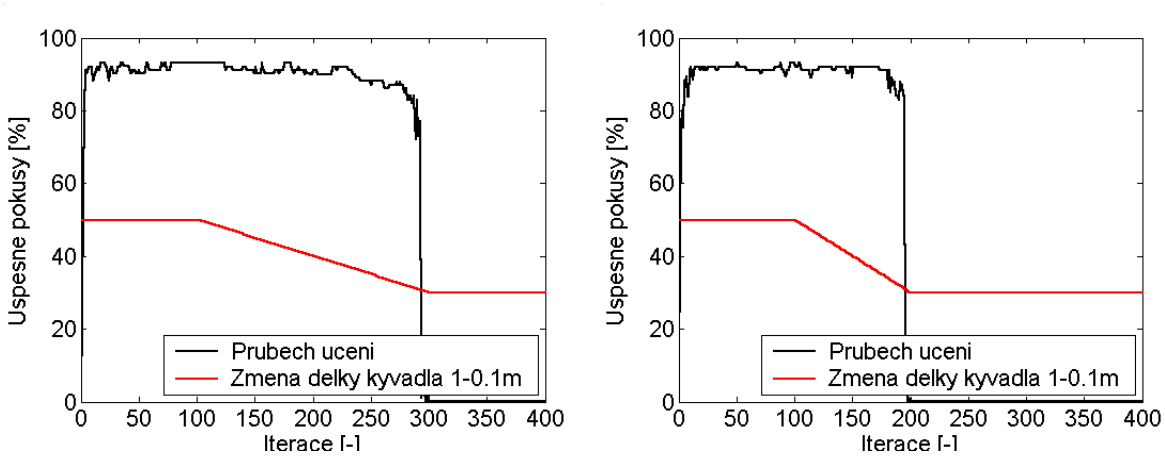
Z obrázků je jasně patrno, že procento úspěšných pokusů se nemění pokud dojde ke změně délky ramena z malého na větší poloměr, jak ukazuje obrázek 2. To je způsobeno snadnějším řízením kyvadla s delším ramenem.

Naopak procento úspěšných pokusů se zhoršuje pokud dojde ke zmenšení délky kyvadla. Kratší kyvadlo je daleko nestabilnější a proto se i hůře hledá výsledný zákon řízení. Na obrázku 3 je velmi dobře vidět jak je procento úspěšných pokusů ovlivněno zmenšením délky kyvadla na 0.6m po 100. iteraci. Řídicí člen je schopen se přizpůsobit této změně během 20ti iterací.



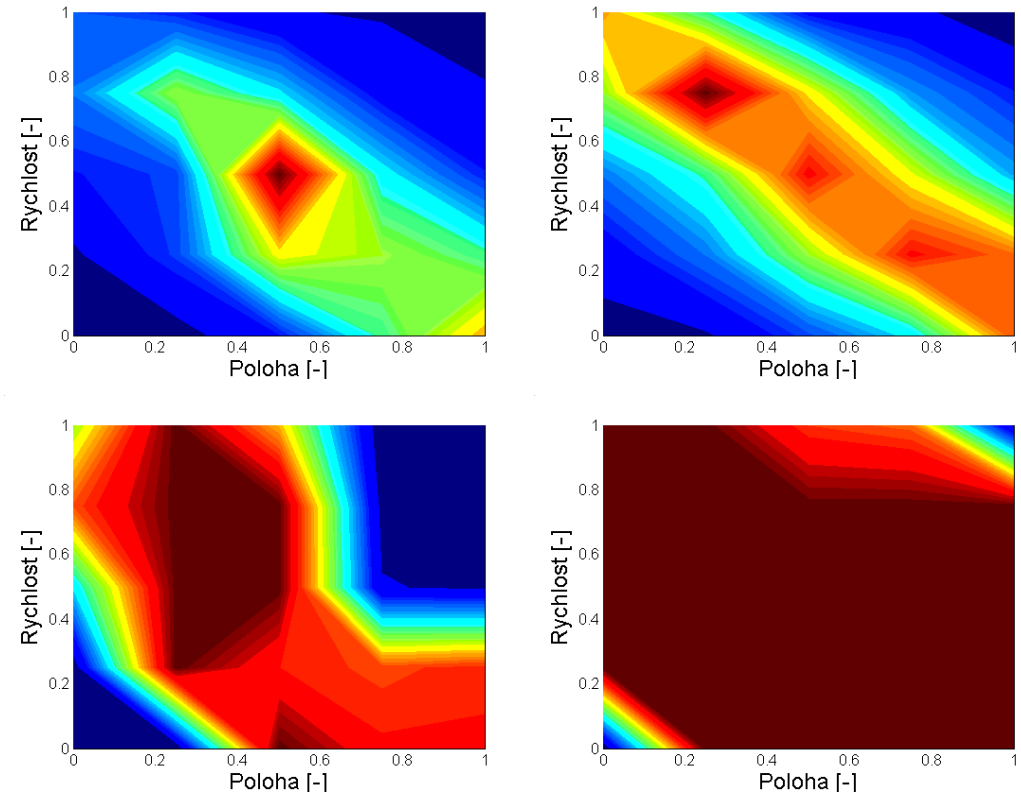
Obr. 5 Vliv zmenšení délky kyvadla z 1 na 0.1m

Obrázky 4 a 5 ukazují jak je procento úspěšných pokusů ovlivněno neúnosně velkou změnou délky kyvadla. Pokud je skoková změna příliž velká dojde k naprosté ztrátě konvergence. Toto je způsobeno tím, že navržený rastr Q-tabulky přestane odpovídat požadovaným vlastnostem kyvadla, jak ukazuje právě obrázek 5.



Obr. 6 Vliv spojité změny délky kyvadla z 1 na 0.1m

V dalších experimentech byl sledován vliv spojité změny délky kyvadla. Během těchto experimentů byla délka kyvadla měněna průběžně z 1 na 0.1m jak ukazuje obrázek 6. Je dobře patrno, že pokud dochází k pomalé změně v parametrech nestacionární soustavy, Q-učení je schopno se velmi dobře přizpůsobit, aniž by docházelo k lokálnímu zhoršení naučení jak bylo ukázáno na obrázku 3.

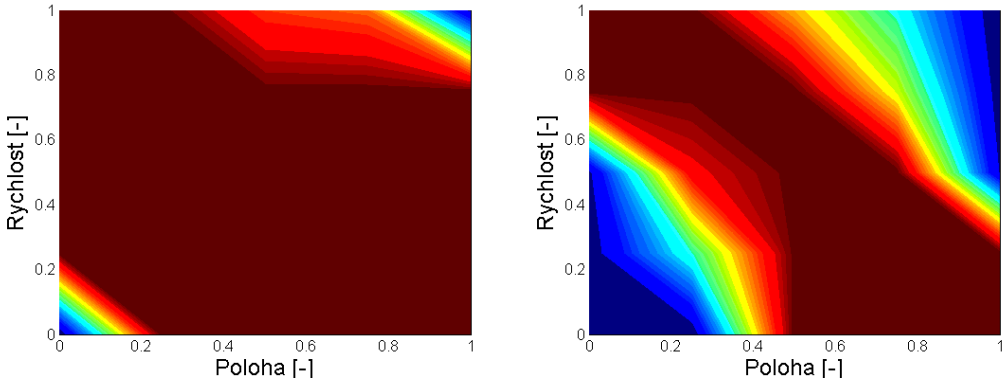


Obr. 7 Vývoj oblasti řiditelnosti během učení

Pro názornost je na obrázku 7 ukázán vývoj oblasti řiditelnosti během učení. Oblast řiditelnosti je vytvořena tak, že na osu *x* je vynesena normalizovaná poloha kyvadla (tedy úhel odklonu kyvadla od vertikální) a na osu *y* je vynesena normalizovná úhlová rychlost kyvadla. Vybraná poloha a rychlost slouží jako počateční podmínky simulací a po provedení určitého ejpočtu pokusů s danými počátečními podmínkami je barvou znázorněno procento uspěšných

pokusů (červená barva odpovída maximálnímu a modrá barva minimálnímu procentu uspěšných pokusů).

Obrázek 8 ukazuje změnu oblasti řiditelnosti po skokové změně délky kyvadla z 1 na 0.6m (průběh učení je zobrazen na obrázku 3).



Obr. 8 Oblast řiditelnosti po skokové změně délky kyvadla z 1 na 0.6m

7. Závěr

V tomto článku jsme se snažily ukázat vlastnosti diskrétní metody Q-učení v souvislosti s navržením řídicího členu pro nestacionární soustavu. Touto soustavou byl jednoduchý simulační model inverzního kyvadla s proměnnou délkou kyvadla. V části praktických experimentů bylo názorně ukázáno jak ovlivňuje velikost změny délky kyvadla výsledné naučení řídicího členu. Praktické experimenty jsou v souladu s teoreticky dokázanými vlastnostmi a ukazují, že je možné Q-učení uspěšně použít při návrhu řídicího členu pro nestacionární soustavu. Z těchto experimentů bude možno vycházet v dalších následných pracích, kdy bude možno nahradit jednoduchý simulační model inverzního kyvadla složitějším modelem aktivního magnetického ložiska a dále o nahrazení diskrétní medoty Q-učení spojitou verzí této metody, která je již také k dispozici.

8. Poděkování

Tato práce vznikla za podpory Ministerstva školství a projektu MSM 0021630518 „Simulační modelování mechatronických soustav“.

9. Literatura

Věchet S., Krejsa J., Březina T.: *Using Q-learning with LWR for Inverted Pendulum Contro*l, Mechatronics, Robotics and Biomechanics 2003, pp.91-92, Hrotovice, 2003