VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

ÚSTAV MECHANIKY TĚLES, MECHATRONIKY A BIOMECHANIKY

Ing. Stanislav Věchet

NÁVRH ROBUSTNÍHO ŘÍZENÍ MODIFIKOVANÝM Q-UČENÍM

THE PROPOSAL OF ROBUST CONTROL THROUGH MODIFIED Q-LEARNING

Zkrácená verze Ph.D. Thesis

Obor:	Inženýrská mechanika
Školitel:	Doc. RNDr. Ing. Tomáš Březina, CSc.
Oponenti:	Prof. Ing. Vladimír Řeřucha, CSc.
	Prof. Ing. Jiří Skalický, CSc.
Datum obhajoby:	20.12.2004

Klíčová slova

Q-učení, Robustní řízení, Aktivní magnetické ložisko

Key Words

Q-Learning, Robust Control, Active Magnetic Bearing

Místo uložení disertační práce

Ústav mechaniky těles, Mechatroniky a Biomechaniky, FSI VUT v Brně

ISBN 80-214-2922-4

ISSN 1213-4198

Obsah

1 Úvod

Návrh inteligentních řídících členů představuje relativně novou oblast použití metod umělé inteligence (Artificial intelligence - AI). Metody AI, zejména ty které používají strojového učení v reálném čase, mohou představovat východiska návrhu nových metod řízení („inteligentních řídicích členů“), případně vyžadujících méně složitou řídicí elektroniku. Učení se dokonce stane zásadním faktorem, jestliže se prostředí nebo cíle subjektu řízení mění v čase.

Disertační práce je zaměřena na aplikaci inteligentních řídicích členů na řízení dynamických soustav. Pomocí metod AI je navrženo robustní řízení, které je v poslední době velmi populární, protože se ukazuje, že v mnoha případech není možné adekvátně popsat regulovanou soustavu odpovídajícím matematickým modelem. Úplný matematický model je možné vytvořit jen pro relativně jednoduché, nebo dostatečně prozkoumané případy. Na základě takového matematického modelu je možné navrhnout optimální regulátor. V ostatních případech, kdy není dostatečně věrně popsána sledovaná soustava, je navržen robustní regulátor na základě zjednodušeného matematického modelu.

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 [37][54]). 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.

Tato práce je zaměřena na rozšíření klasického Q-učení, které pracuje pouze s diskrétními stavy soustavy, o možnost práce se spojitými stavy, tak jak jsou prezentovány reálnou soustavou. Jednou z možností, jak pracovat se spojitým prostředím, je použití standardního stavového prostoru, který je vytvořen diskretizací spojitého prostředí. Již volba vhodné diskretizace může být problematická a v řadě případů vede ke ztrátě konvergence [25][27][34]. V případě velkého množství stavových proměnných je diskretizace velmi neefektivní a přispívá k velké výpočetní náročnosti. Jinou možností je použít pro reprezentování stavů prostředí vhodný aproximátor.

Praktické experimenty jsou prováděny jednak na jednoduchém simulačním modelu inverzního kyvadla, a zejména na složitějším simulačním modelu aktivního magnetického ložiska, k němuž je k dispozici i experimentální laboratorní model.

Tato práce je prováděna ve spolupráci s řešitelským týmem Laboratoře mechatroniky a robotiky Ústavu mechaniky těles VUT v Brně.

2 Formulace problému a cílů jeho řešení

Problematika návrhu „inteligentního řídicího členu“ představuje moderní oblast výzkumu. Prvním úkolem bylo dostatečně vymezit předmět zkoumání a definovat požadované cíle práce.

Pro návrh inteligentních řídících členů existuje řada metod. Velmi perspektivní se jeví diskrétní metoda Q-učení [1,2,53,62,63]. S diskrétní formou této metody bylo dosaženo dobrých výsledků také v Laboratoři mechatroniky a robotiky[1]. Vzhledem k tomu že řízené veličiny i veličiny popisující stavy řízené soustavy jsou u reálných soustav prakticky vždy spojité, je logickým krokem metodu Q-učení rozšířit právě o možnost pracovat se spojitými veličinami, což je cílem této práce..

Cílem práce je tedy analyzovat možnost rozšíření klasického modelu Q-učení, pracujícího s diskretizovanými veličinami, o schopnost využívat spojité veličiny stavů simulované soustavy, a to s využitím dostupných aproximátorů. Dále pak nalézt veličiny podstatně ovlivňující chování takto modifikovaného učení, případně nalézt optimální metodu volby těchto parametrů z hlediska aplikace na různé modelové soustavy.

Dílčí cíle lze formulovat takto:

- analyzovat současný stav poznání o použití Q-učení pro řízení spojitými veličinami
- vybrat a implementovat vhodný aproximátor
- modifikovat klasický algoritmus Q-učení pomocí vhodného aproximátoru, tak aby umožňoval řízení pomocí spojitých veličin
- na úlohách simulačního modelování porovnat vlastnosti klasického a modifikovaného algoritmu Q-učení

[1] jedná se především o habilitační práci Doc. Březiny a související publikace řešitelského týmu Laboratoře mechatroniky a robotiky.

3 Přehled současného stavu řešené problematiky

3.1 Metoda opakovaně posilovaného učení

Metody RL řeší problém nalezení vzorů požadovaného chování řídicího členu jako optimalizační úlohu, která použitím jednoduchého okamžitého ohodnocování zásahu subjektu do prostředí postupně zlepšuje odhad výkonu tohoto subjektu. Na základě dosaženého odhadu celkového výkonu se potom vybírají optimální zásahy do prostředí tak, aby tyto zásahy celkový výkon extremalizovaly. Ohodnocování očekávaného výkonu subjektu může zahrnovat nejrůznější kritéria, jako např. minimální čas, minimální cenu, minimum kolizí. Probíhá-li proces učení v reálném čase, může být subjekt chápán jako řídicí člen a prostředí jako řízená soustava. Základním rysem opakovaně posilovaného učení je tak schopnost v průběhu procesu učení zlepšovat chování řízené soustavy.

3.2 Metoda Q-učení

Jak již bylo zmíněno v předcházejících odstavcích, je metoda Q-učení variantou opakovaně posilovaného učení. Při tvoření celkového obrazu o principu této metody i o její praktické aplikaci bylo ve velké míře použito článků a literatury, dostupné přes internet. Řada zajímavých prací byla k dispozici pouze jako internetové stránky a jako takové nebyly nikdy otištěny v klasické papírové podobě.

Velká většina prací, zabývajících se Q-učením, se zabývá teoretickým rozpracováním formalismu metody a důkazy konvergence[56][58]. Další skupina publikací je tvořena popisy simulačních experimentů, které prověřují teoreticky dokázané vlastnosti metody[32][49][50]. Poměrně malou, avšak zajímavou skupinu publikací, tvoří teoretické rozpracování metod, které umožňují v jisté míře měnit vlastnosti Q-učení tak, jak bylo původně navrženo (Watkins[63][64]). Patrně nejméně je zastoupena oblast zabývající se praktickou aplikací Q-učení na řízení reálných soustav.

V předložené disertační práci se zabýváme možností rozšířit diskrétní metodu Q-učení na možnost pracovat přímo se spojitými veličinami soustavy, bez nutnosti diskretizace. Velmi zajímavá teoretická studie zabývající se tzv. „spojitým“ Q-učením je

- **R.Munos:** ***A convergent Reinforcement Learning algorithm in the continuous case based on a Finite Difference method*** **[44]** - Tato práce se v hlavní míře zabývá zajímavou metodou kontinualizace Q-učení založenou na metodě konečných diferencí resp. konečných prvků (Munos[43])

4 Teoretický rozbor použitých metod

4.1 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 $r(s_t, s_{t+1}, a_t)$, která určuje konkrétní odměnu/pokutu za přechod ze stavu s_t do stavu s_{t+1} při použití akce a_t. Dlouhodobý cíl agenta je definován jako funkce okamžitých odměn, například kumulativní srážková odměna (cumulative discount reward) $\sum_{t=0}^{\infty} \gamma^t r(s_t, s_{t+1}, a_t)$, kde $0 \le \gamma < 1$ je srážkový faktor, řídící 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 očekávanou kumulativní srážkovou odměnu při počátečním stavu s a použití strategie π.

4.2 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 následném pokračování v dané strategii. Konkrétní hodnotou hodnotové funkce je potom maximum Q-hodnot pro daný stav $f(s) = \max_a Q(s,a)$. Q-funkce může být implementována různými způsoby, v použitém případě implementací tabulkou je přepočtový vztah pro Q-funkci:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left[r(s_t, a_t, s_{t+1}) + \gamma \max_{a_t} Q(s_{t+1}, a_t) - Q(s_t, a_t) \right] \qquad (4.2.1)$$

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 a akcí). Q-hodnoty konvergují s pravděpodobností jedna v případě, že jsou v průběhu učení všechna uzlová místa navštívena nekonečně krát (každá akce je v každém stavu vykonána nekonečně krát během nekonečného množství kroků), konvergence tedy může být značně pomalá.

4.3 Použití spojitých stavů a akcí v Q-učení

Klasické Q-učení pracuje s diskrétní Q-funkcí, nejčastěji implementovanou jako tabulka Q-hodnot. Jako indexy této tabulky jsou používány diskrétní hodnoty stavů a akcí. Při tomto způsobu implementace Q-funkce je také důležité správné rozdělení mřížky tabulky.

Pokud je ovšem potřeba pracovat se soustavou s mnoha stupni volnosti a pro adekvátní model této soustavy je zapotřebí pracovat s větším počtem stavových proměnných, popř. pracovat se spojitými stavy a akcemi, je výhodné najít jinou možnost implementace Q-funkce, než nabízí klasická tabulka Q-hodnot. Jako možnost této implementace se jeví nahrazení tabulky vhodným aproximátorem. Jako vhodný aproximátor byl zvolen algoritmus patřící do skupiny paměťově orientovaných, také nazývaných *Lazy Learning,* s lokálním modelem (Locally Weighted Learning - LWL).

4.4 Lokálně vážené učení

Metody *Lazy Learning* odkládají zpracování trénovacích dat do chvíle, kdy je potřeba zodpovědět dotaz. Metoda obvykle vyžaduje ukládání trénovacích dat do paměti a pro zodpovězení konkrétního dotazu hledání odpovídajících dat v databázi. Tento typ učení je často popisován jako paměťově orientované (*memory-based*) učení. Jedna z forem *Lazy Learning* hledá množinu bodů nejbližšího okolí a z těchto bodů je následně tvořena predikce, tato forma nese název Lokálně vážené učení (*Locally Weighted Learning*) a používá lokální model.

4.4.1 Lokálně vážená regrese

Lokálně vážená regrese je založena na metodě nejmenších čtverců, avšak metoda LWR bere v úvahu i vzájemnou polohu bodů, to může být dosaženo např. vážením dat. V tomto případě není regresní přímka konstruována na základě celého datového souboru, ale pouze z bodů v určité definované oblasti a závisí na vzájemné poloze jednotlivých bodů dané oblasti. Parametry regresní přímky jsou přímo závislé na vzdálenosti od odhadovaného bodu q. Nejčastějším kritériem vzdálenosti je Euklidovská vzdálenost $d_E([x_i;y_i],\mathbf{q})$. V tomto jednoduchém ilustračním případě je vzdálenost daná pouze jednoduchým vztahem $d_E\left(\left[x_i, y_i\right],\mathbf{q}\right)=\sqrt{\left(x_i - x_q\right)^2}$. Na základě takto získané vzdálenosti je poté vypočítána váha každého bodu. Jako nejvhodnější funkce pro výpočet váhy daného bodu jsou označovány kernelové funkce hladké, symetrické, zvonového tvaru. Nejčastěji používanou funkcí bývá funkce založena na gausově jádru:

$$K\left(d_E\right)=e^{-d_E^2} \qquad (4.4.1.1)$$

Postup výpočtu metodou LWR je následující. Nejdříve je třeba vypočítat vzdálenost a váhu jednotlivých bodů $w_i = \sqrt{K\left(d_E\left(\left[x_i, y_i\right], \mathbf{q}\right)\right)}$. Pro další výpočty je nutné tyto váhy umístit do diagonální matice **W,** kde jednotlivé prvky na digonále odpovídají jednotlivým vypočítaným vahám $W_{ii} = w_i$. Na základě takto vypočítané matice vah přepočítáme známou matici vstupů **X**, získáme tak váženou matici vstupů **Z** = **W X**. Vlastní výpočet neznámých parametrů lze poté zapsat jako:

$$\begin{aligned} \mathbf{Z}^T\mathbf{y} &= \left(\mathbf{Z}^T\mathbf{X}\right)\boldsymbol{\beta} \\ \boldsymbol{\beta} &= \left(\mathbf{Z}^T\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{y} \end{aligned} \qquad (4.4.1.2)$$

Takto vypočítanou matici parametrů $\boldsymbol{\beta} = \left[p_1 \quad p_2\right]^T$ lze použít pro proložení regresní přímky, případně odhadovat jednotlivě neznáme body.

5 Experimentální modely

Jako experimentálních modelů pro všechny experimenty bylo použito jednoduchého simulačního modelu inverzního kyvadla a jednohmotového aktivního magnetického ložiska. Pro ověření teoreticky získaných poznatků v reálném prostředí byl použit fyzikální model aktivního magnetického ložiska, který byl vyroben na Ústavu výkonové elektrotechniky a elektroniky FEKT v Brně.

5.1 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 vlastnostem nového přístupu v Q-učení. Dále byl použit z důvodu, že disponuje stejnou nestabilitou jako podstatně složitější model aktivního magnetického ložiska.. Na tomto jednoduchém modelu byly prováděny počáteční experimenty se spojitým Q-učením. Použitý simulační model inverzního kyvadla lze popsat rovnicemi:

$$\begin{aligned} &\left(M+m\right)\ddot{x} + ml\ddot{\varphi}\cos\varphi - ml\dot{\varphi}^2\sin\varphi = F \\ &\frac{12}{13l}\left(g\sin\varphi + \ddot{x}\cos\varphi\right) = \ddot{\varphi} \end{aligned} \qquad (5.1.1)$$

, kde M je hmotnost vozíku, m homotnost 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

5.2 Simulační model aktivního magnetického ložiska

Použitý simulační model AMB je jednohmotový a sestává z rotoru a dvou elektromagnetů, které zajišťují v každém časovém okamžiku správnou polohu rotoru v dané ose. Pro praktické užití je ovšem zapotřebí kontrolovat polohu rotoru ve dvou

osách a to horizontálně i vertikálně. Poloha rotoru je snímána pomocí bezdotykových snímačů a na základě těchto signálů jsou pomocí řídicího členu kontrolovány napájecí proudy elektromagnetů ložiska. Simulační výsledky pro jednu osu lze použít pro návrh dvou nezávislých regulačních členů pro každou osu. Přitažlivá síla elektromagnetu je popsána obecným vztahem

$$F = K_m \frac{I_m^2}{(x+a)^2} \tag{5.2.1}$$

kde F je přitažlivá síla elektromagnetu, I_m je proud vinutím, x je velikost vzduchové mezery, a je konstanta určující sílu F při $x=0$, K_m je konstanta závisející na síle, nutné ke zvednutí vodorovně umístěného rotoru. Linearizace chování elektromagnetu je provedena následujícím způsobem. Je-li velikost proudu elektromagnetu I_m vypočtena z požadované velikosti regulačního zásahu I vztahem $I_m = \sqrt{I}(x+a)$, kde x je skutečná velikost vzduchová mezery určená snímačem, pak prosílu F dostáváme:

$$F = K_m \frac{\left(\sqrt{I}(x+a)\right)^2}{(x+a)^2} = KI \tag{5.2.2}$$

tedy přitažlivá síla je úměrná požadovanému regulačnímu zásahu a nezávisí na skutečné velikosti mezery. Vztah mezi silou elektromagnetu a polohou hmotného rotoru je dán diferenciální rovnicí 2. řádu $\ddot{x}m = F + F_e(t)$, kde m je hmotnost rotoru, $\ddot{x}$ je zrychlení rotoru, $F_e(t)$ je zátěžná síla.

Předchozí rovnice jsou převzaty z [36] a na jejich základě byl sestaven simulační model pro Matlab. Pro další experimenty byl tento model kompletně přepsán tak aby bylo možno jej používat v programovacím jazyce Delphi.

6 Provedené experimenty a prezentace výsledků

6.1 Okrajové podmínky experimentů

Soustavy byly obecně testovány 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 resp. ložiska v určeném intervalu. Pokud toho bylo dosaženo, byl pokus vyhodnocen jako úspěšný.

6.1.1 Výpočtové modely

K simulacím popsaným v této části byly použity výpočtové modely, které byly podrobněji popsány v kapitole 5. V následujících odstavcích se budeme podrobněji zabývat pouze upřesněním použitých parametrů modelů a podmínek, za jakých byly tyto modely použity. V experimentech tedy bylo pro jednotlivé soustavy použito následujících parametrů modelů:

Inverzní kyvadlo: $M = 0.2\left[kg\right], m = 0.1\left[kg\right], g = 9.81\left[kgms^{-2}\right], l = 0.5\left[m\right],$

AMB: $R = 266\left[\Omega\right], L = 0.87\left[H\right], m = 0.2\left[kg\right], K_m = 1.37\times10^{-6}\left[-\right], a = 1.43\times10^{-3}\left[-\right]$

Tyto parametry byly použity za předpokladů, že pro inverzní kyvadlo nebyly uvažovány pasivní odpory. Pro AMB byl symetrický tuhý rotor nahrazen hmotným bodem, zanedbána vazba mezi vodorovným a svislým kmitáním, tj. kmitání pouze v jedné rovině a byl zanedbán vliv tíže.

6.1.2 Simulační přístupy

Simulační experimenty byly hodnoceny následujícím způsobem. Pro hodnocení výsledků simulací byl zaved termín *pokus*. Provedení 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 řídicích 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 jednotlivých modelů.

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.

V modelu AMB byla v každém pokusu modelována také zátěžná síla $F_e(t)$. Její průběh odpovídal schodovité funkci s náhodnou velikostí konstantních částí z intervalu $\langle -10, 10 \rangle [\mathrm{N}]$. Aby byla zajištěna srovnatelnost testů, byly pro zmíněných 100 pokusů vygenerovány:

- náhodné počáteční podmínky, a pro každý pokus různý průběh zátěžné síly, takto vygenerovaná data byla uložena do souboru a označena jako *data A*
- počáteční podmínky, které byly generovány takovým způsobem, aby rovnoměrně pokryly stavový prostor. Pro tyto pokusy byl vygenerován jednotný průběh zátěžné síly. Takto připravená data byla uložena a označena jako *data B*.

Data A byla použita pro veškeré testy, tj pro:

- stanovení průměrných délek pokusu během fáze učení,
- porovnání chování jednotlivých modelů řízených strategií dosaženou pomocí diskrétního Q-učení oproti spojitému Q-učení.

Data B byla použita pro srovnání *oblastí řiditelnosti*, tedy srovnání počátečních podmínek, pro které je soustava řiditelná.

Poznamenejme, že při použité organizaci simulací není dosažitelná průměrná úspěšnost pokusů 100 %. Je to dáno poměrně hrubým odhadem množiny řiditelných stavů, jejíž prvky byly používány jako počáteční stavy jednotlivých pokusů. Na druhé straně tento odhad umožňuje lépe porovnat úspěšnost získaných strategií pro diskrétní a spojité Q-učení.

6.1.3 Standardní podmínky simulací

Simulace byly prováděny ve dvou etapách. První etapa byla vyhledávací a sloužila k odhadu toho, jak konkrétně závisí rychlost učení a kritérium kvality řízení na konkrétních volbách hodnot parametrů simulací (*podmínky simulací*). Její dílčí výsledky nebyly do této práce zahrnuty. Z hodnot parametrů jednotlivých simulací, pro které bylo dosaženo nejlepších výsledků, byly sestaveny *standardní podmínky simulace*. Jsou shrnuty v tabulkách 6.1.3.1 a 6.1.3.2, kde jsou zobrazeny hodnoty parametrů pro diskrétní i spojité Q-učení. V druhé etapě simulací, jejíž výsledky již jsou uvedeny v této práci, bylo provedeno ověření získaných hodnot parametrů. V jednotlivých sadách simulací byly brány jako výchozí standardní podmínky simulace, vůči kterým byla měněna typicky hodnota jednoho parametru.

	Inverzní kyvadlo	AMB
Množina řídicích akcí	$\{-a_{\max}, 0, a_{\max}\}$	$\{-u_{\max}, 0, u_{\max}\}$
Okamžité posílení	$r = \begin{cases} 1 & \varphi \le \lvert\varphi_{\min}\rvert \\ 0 & \varphi \in (\lvert\varphi_{\min}\rvert, \lvert\varphi_{\max}\rvert) \\ -1 & jinak \end{cases}$	$r = \begin{cases} 1 & x \le \lvert x_{\min}\rvert \\ 0 & x \in (\lvert x_{\min}\rvert, \lvert x_{\max}\rvert) \\ -1 & jinak \end{cases}$
Parametry učení	$\alpha_0 = 0.2,\ \gamma = 0.99$	$\alpha_0 = 0.2,\ \gamma = 0.99$
Zátěžná síla	nepoužita	náhodná schodovitá, $(F_e)_{max} = 10[\mathrm{N}]$

Tab. 6.1.3.1: Použité parametry diskrétního Q-učení

	Inverzní kyvadlo	AMB
Interval řídicích akcí	$\langle -a_{\max}, a_{\max} \rangle$	$\langle -u_{\max}, u_{\max} \rangle$
Okamžité posílení	$r = \begin{cases} 1 & \varphi \le \lvert\varphi_{\min}\rvert \\ 0 & \varphi \in (\lvert\varphi_{\min}\rvert, \lvert\varphi_{\max}\rvert) \\ -1 & jinak \end{cases}$	$r = \begin{cases} 1 & x \le \lvert x_{\min}\rvert \\ 0 & x \in (\lvert x_{\min}\rvert, \lvert x_{\max}\rvert) \\ -1 & jinak \end{cases}$
Parametry učení	$\alpha_0 = 0.2,\ \gamma = 0.99$	$\alpha_0 = 0.2,\ \gamma = 0.99$
Zátěžná síla	nepoužita	náhodná schodovitá, $(F_e)_{max} = 10[\mathrm{N}]$

Tab. 6.1.3.2: Použité parametry spojitého Q-učení

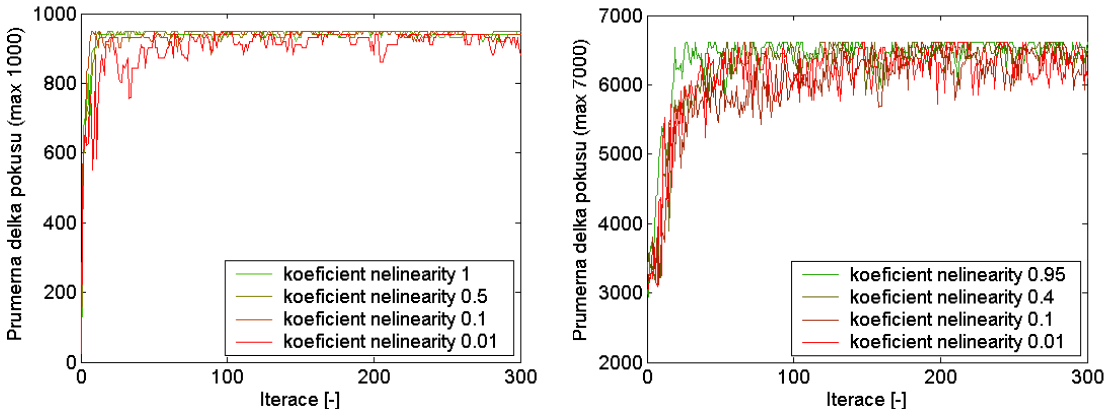
6.2 Diskrétní Q-učení

6.2.1 Volba rastru tabulky

V algoritmu diskrétního Q-učení jsou spojité hodnoty stavů a akcí diskretizovány. Pomocí takto diskretizovaných veličin je vytvořena diskrétní tabulka Q-hodnot. Tabulka je tvořena sousedícími *buňkami*. *Buňka* je definována jako *n* rozměrná oblast v *n* rozměrném prostoru stavů a akcí. Rozměry buňky odpovídají dělení rastru tabulky. Není nutné, aby rozdělení rastru bylo lineární. Právě vhodně zvolená nelinearita rastru umožňuje ve většině případů výrazně ovlivnit kvalitu naučení. Je totiž možné zvolit hustější síť v okolí důležité oblasti, jakou je například požadovaná poloha, oproti méně husté síti v méně důležitých okrajových oblastech.

Akce byla vždy dělena do tří buněk pro tři hodnoty akcí z množiny $\{-a_{\max}, 0, a_{\max}\}$ pro inverzní kyvadlo, resp. $\{-u_{\max}, 0, u_{\max}\}$ pro aktivní magnetické ložisko.

Na obrázku 6.2.1.1 je znázorněn průběh učení pro jednotlivé nelinearity rastru, různé počty buněk a oba použité modely.

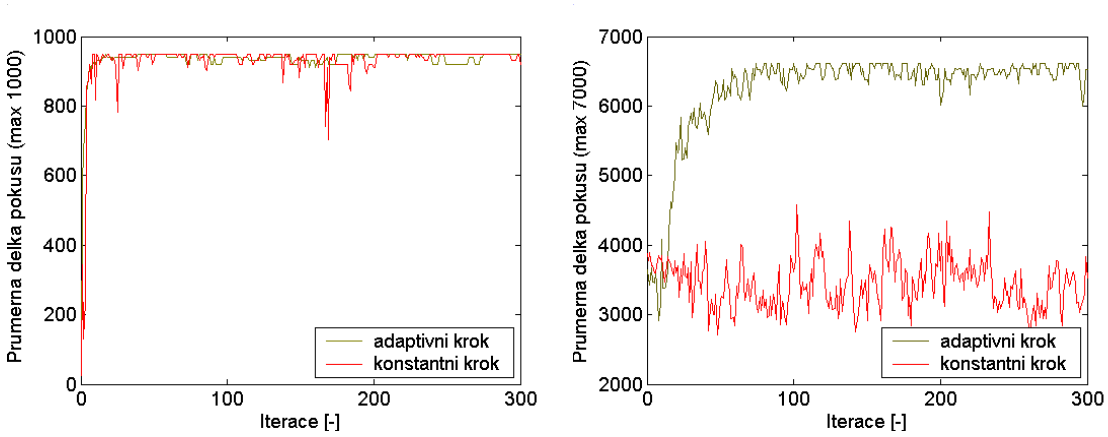


Obr. 6.2.1.1: Vliv nelinearity rastru na inverzní kyvadlo(vlevo) a AMB(vpravo)

6.2.2 Adaptivní integrační krok

Proces učení probíhá v závislosti na přechodech stavů mezi jednotlivými buňkami. V čase t je vybrán pár stav-akce (s_t, a_t) z buňky D_t (i_t, j_t), po provedení jednoho časového kroku simulace se soustava přesune do nového stavu s_{t+1}. Pro proces učení je důležité aby se nový stav nacházel v jiné, nejlépe v sousední buňce tabulky (pokud konkrétní hodnoty veličiny odpovídají stejným indexům rastru, není mezi nimi při Q-učení rozdíl). Při použití konstantního časového kroku je tato podmínka splněna jen v malém počtu případů.

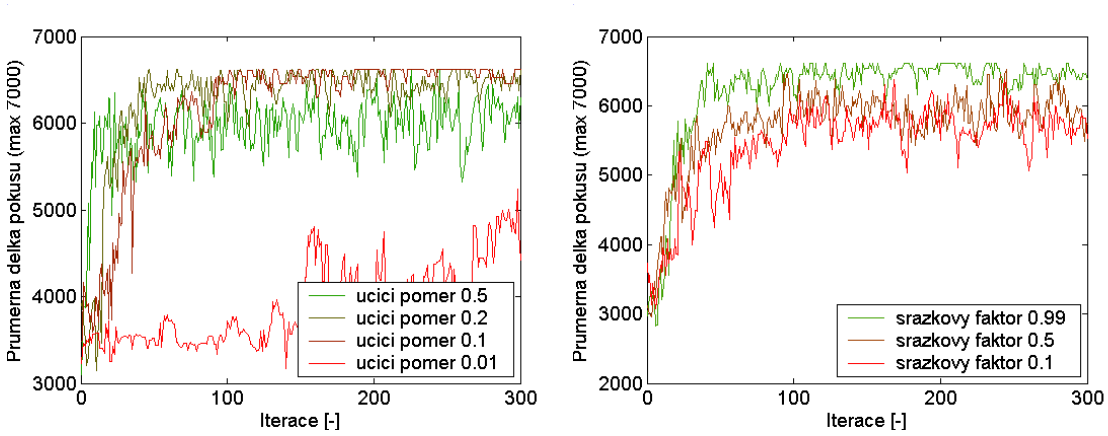
Jako velmi výhodné se ukázalo použití tzv. *adaptivního integračního kroku*, kdy je aktuální časový krok τ prodlužován tak dlouho, dokud soustava nepřejde do sousední buňky, případně dosáhne nekorektního stavu. Časový interval byl prodlužován tak, aby nepřekročil délku trvání 100×10^{-3}[s] pro model inverzního kyvadla a 1×10^{-3}[s] pro model AMB, tj. časový krok se pohyboval v intervalu $\langle 10\times10^{-3}, 100\times10^{-3}\rangle$[s] resp. $\langle 10\times10^{-6}, 1\times10^{-3}\rangle$[s]. Vliv adaptivního časového kroku je jasně patrný z obrázku 6.2.2.1.



Obr. 6.2.2.1: Vliv adaptivního integračního kroku; Inverzní kyvadlo(vlevo) a AMB(vpravo)

6.2.3 Parametry učení

Další vliv na kvalitu učení mají parametry α a γ, tedy učící poměr a srážkový faktor. Následující obrázky 6.2.3.1. ukazují vliv těchto parametrů na kvalitu naučení. Je vidět že vliv nastavení parametru γ se zvyšuje se snižováním počtu buněk tabulky. Oproti tomu lze říci, že velikost tabulky nemění velikost vlivu nastavení parametru α na učení.



Obr. 6.2.3.1: AMB; Vliv učícího poměru a srážkového faktoru

6.2.4 Volba posilovací funkce

Jak bylo ukázáno v [28], nemá posilovací funkce výrazný vliv na průběh učení. Proto bylo jako posilovací funkce použito pouze tzv. *jednoduché posílení*. Posilovací funkce stanovuje odměnu/trest pro jednotlivé simulační modely

z množiny $\{-1,0,1\}$ pravidlem $r=\begin{Bmatrix} 1 & \varphi \le |\varphi_{\min}| \\ 0 & \varphi \in (|\varphi_{\min}|,|\varphi_{\max}|) \\ -1 & jinak \end{Bmatrix}$ pro inverzní kyvadlo a pravidlem $r=\begin{Bmatrix} 1 & x \le |x_{\min}| \\ 0 & x \in (|x_{\min}|,|x_{\max}|) \\ -1 & jinak \end{Bmatrix}$ pro aktivní magnetické ložisko

6.3 Spojité Q-učení

6.3.1 Velikost spojitého prostoru

Jednotlivé osy prostoru tedy nejsou rozděleny na konečný počet intervalů ale jsou na ně vynášeny spojité hodnoty. Aby bylo možné provozovat spojité Q-učení, bylo nutné nejprve normalizovat všechny použité stavové a akční veličiny.

Nejprve jsou tedy získány reálné hodnoty z modelu, tyto jsou poté normalizovány a s takto upravenými veličinami je prováděn vlastní algoritmus Q-učení. Pokud nejsou hodnoty normalizovány není možné jednoduše posuzovat vzdálenosti jednotlivých bodů v prostoru, což způsobuje komplikace při použití aproximátoru.

6.3.2 Volba parametrů LWR

Jak bylo řečeno v předcházejících kapitolách, ve spojitém Q-učení je diskrétní tabulka nahrazena spojitým prostorem a vhodným aproximátorem. V našem případě byla jako aproximátor zvolena metoda LWR. Standardní parametry LWR byly nastaveny podle tabulky 6.3.2.1, všechny parametry jsou vztaženy k normalizovanému prostoru a jsou tedy bezrozměrné.

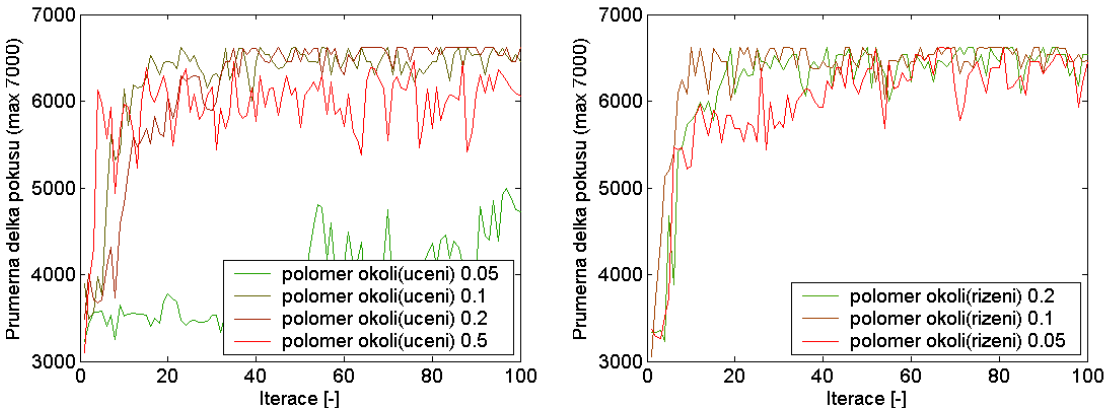
poloměr okolí (učení)	r_l	0.2
poloměr okolí (řízení)	r_c	0.1
šířka pásma	h	0.1
šířka základny	D	500
minimální počet bodů	k_{min}	10
koeficient aktualizace	c_a	1

Tab. 6.3.2.1: Parametry LWR

Během simulací bylo zjištěno, že poloměr okolí r, ze kterého je počítán odhad, je dobré volit jako různě veliký pro fázi učení a pro fázi řízení. Ukázalo se, že optimální hodnota parametru r_c je polovina hodnoty parametru r_l. Je to do značné

míry dáno nastavením parametru D, který během fáze učení také omezuje vliv okolních bodů na aproximaci.

Následující graf 6.3.2.1 zobrazuje vliv vybraných parametrů na kvalitu učení. Grafy ukazují průběh učení pro 100 provedených iterací, protože dále se již kvalita naučení výrazně nemění. Výpočetní náročnost je omezena především parametry r_l a h. Je to dáno tím, že na hodnotách těchto parametrů přímo závisí počet bodů, který bude použit pro aproximaci.



Obr. 6.3.2.1: AMB; vliv poloměru okolí při učení(vpravo) a řízení(vlevo)

6.3.3 Adaptivní integrační krok

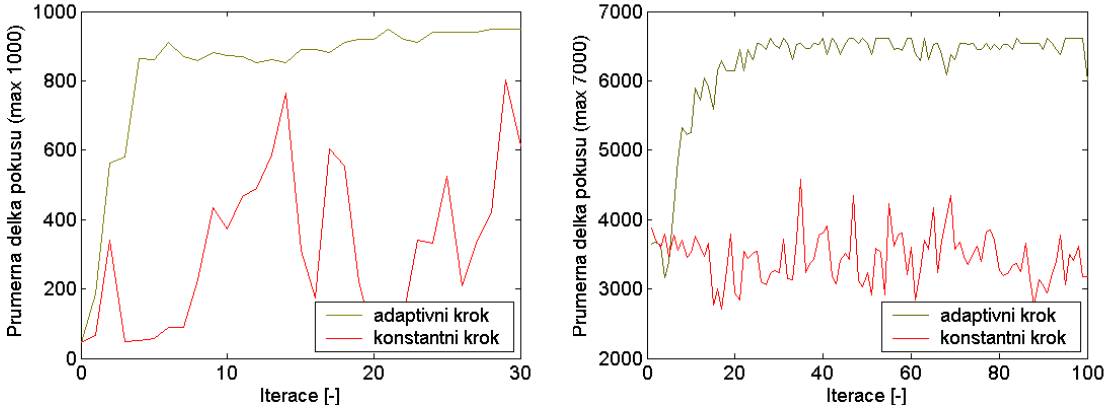
V počátečních experimentech se spojitým učením se jevilo použití adaptivního kroku jako bezvýznamné. Jak ovšem ukázaly následující experimenty, použití adaptivního integračního kroku má významný vliv na kvalitu naučení.

Proces učení probíhá v závislosti na přechodech mezi různými stavy. V čase t je tedy vybrán pár stav-akce (s_t, a_t) ze spojitého prostoru. Po provedení jednoho časového kroku simulace se soustava přesune do nového stavu s_{t+1}. Jelikož se nacházíme ve spojitém prostoru je vysoce pravděpodobné že $s_t <> s_{t+1}$ Proto se původní idea neimplementovat adaptivní integrační krok zdála jako opodstatněná. Jak bylo ovšem zjištěno dalšími simulacemi, pro proces učení je dobré, aby se nový stav nacházel v dostatečné vzdálenosti od bodu původního. Při použití konstantního časového kroku je tato podmínka splněna jen v malém počtu případů. Jako velmi výhodné se tedy ukázalo adaptivní integrační krok opět implementovat.

Aktuální časový krok τ je prodlužován tak dlouho dokud soustava nepřejde do stavu dostatečně vzdáleného od stavu původního, případně dosáhne nekorektního stavu. Minimální vzdálenost od původního stavu je značena r_a a vliv velikosti tohoto parametru je zobrazen na obrázku 6.3.3.1. Časový interval byl prodlužován tak, aby nepřekročil délku trvání 100×10^{-3}[s] pro model inverzního kyvadla a 1×10^{-3}[s] pro model AMB, tj. časový krok se pohyboval v intervalu $\langle 10\times10^{-3}, 100\times10^{-3}\rangle$[s] resp.

$\langle 10\times10^{-6}, 1\times10^{-3}\rangle$[s]. Použití této strategie během učení výrazně přispělo ke zlepšení konvergence a rychlost učení.

Obrázek 6.3.3.1 ukazuje vliv adaptivního časového kroku a velikosti adaptivního okolí určeného parametrem r_a.



Obr. 6.3.3.1: Vliv adaptivního integračního kroku, Inverzní kyvadlo(vpravo), AMB (vlevo)

7 Porovnání diskrétního a spojitého Q-učení

Na základě simulačních výsledků, popsaných v předchozích kapitolách byly sestaveny *standardní podmínky simulací*, uvedené v kapitole 6. Tyto podmínky byly doplněny o další parametry, které jsou nezbytné pro úspěšné použití popsaných algoritmů.

Nastavení parametrů metody LWR pro oba použité modely se liší jen velmi málo. To je dáno ve velké míře tím, že jsou dané vstupní a výstupní veličiny soustavy normalizovány. Jediný parametr, který je potřeba posuzovat v závislosti na použitém modelu, je parametr r_a. Je to dáno odlišným chováním modelů během delšího časového kroku simulace. Stabilnější model je možné nechat neřízený po delší časový interval a je tedy možné, aby dosáhl během tohoto časového intervalu větší vzdálenosti od původního stavu, aniž by se dostal do stavu nekorektního pro danou soustavu.

Z experimentů, uvedených v kapitole 6, je jasně vidět, že z hlediska průměrné délky pokusu jsou algoritmy diskrétního a spojitého Q-učení rovnocené. Ovšem při použití spojitého stavového prostoru je vidět značně rychlejší konvergence během počátečních fází učení. Je tedy možné pomocí spojitého Q-učení výrazně zkrátit počet iterací nutných pro naučení.

8 Závěr

Předložená disertační práce je zaměřena na rozšíření klasického algoritmu Q-učení, pracujícího s diskrétními množinami stavů a akcí, o možnost práce ve spojitém prostoru. V práci je diskutována metoda kontinualizace Q-učení pomocí lokálních aproximátorů, konkrétně metody Lokálně vážené regrese (Locally Weighted Learning – LWR; kapitola 4.4).

Cíle disertační práce byly splněny ve stanoveném rozsahu a je možné jej shrnout v následujících bodech:

1. První fáze byla věnována průzkumu související literatury a vyhledávání perspektivních metod vhodných ke kontinualizaci Q-učení. Stručný přehled souvisejících témat je uveden v kapitole 3.
2. Další práce byly zaměřeny na výběr a implementaci vhodného aproximátoru. Jako vhodný aproximátor byla zvolena metoda LWR a pro počáteční experimenty implementována v systému MATLAB.
3. Souběžně s výběrem vhodného aproximátoru probíhala implementace algoritmu diskrétního Q-učení v prostředí Delphi 7 a jeho následná kontinualizace s využitím metody LWR.
4. Hlavní částí práce bylo provedení numerických experimentů.
   - Během počátečních fází testování byly hledány standardní podmínky simulací (kapitola 6.1.3) a od těchto podmínek se poté odvíjela hlavní fáze simulací zaměřená na identifikování vlivu jednotlivých parametrů na daný diskrétní resp. spojitý algoritmus Q-učení. U diskrétního Q-učení byl hledán např. správný způsob diskretizace stavových veličin. U spojitého Q-učení byly hledány parametry související s metodou LWR, které mají podstatný vliv na výsledné naučení.
   - V rámci hledání optimálních parametrů spojitého Q-učení bylo zjištěno, že je možné nastavení parametrů aproximátoru, ve spojitosti se spojitým stavovým prostorem, nezávisle na metodě Q-učení, což v konečném důsledku umožňuje používat stejné nastavení pro různé simulační modely (kapitola 7).
   - Hlavním přínosem spojitého Q-učení je odstranění problémů s diskretizací stavového prostoru. Díky možnosti použít nastavení parametrů aproximátoru nezávisle na použitém modelu je výrazně zjednodušeno použití této metody pro různé simulační modely. Jak bylo také ukázáno v kapitole 6.3, bylo dosaženo u spojitého Q-učení vyšší rychlosti učení, při zachování kvality řízení, což v konečném důsledku zkracuje dobu nutnou pro testování vlastností daného modelu.

Zkoumanou problematiku nelze v žádném případně považovat za uzavřenou. Ačkoli bylo ukázáno, že je možné do značné míry zobecnit nastavení parametrů nezávisle na použitém modelu, je rozdíl mezi algoritmy spojitého a diskrétního Q-učení značný. Vlastnosti byly testovány pouze na dvou simulačních modelech, i když značně rozdílných. Pro další porovnání vlastností obou algoritmů je tedy možné pokračovat s ověřovacími simulacemi na dalších simulačních modelech. Je také otevřena možnost nahradit stávající aproximační metodu LWR metodou propracovanější, jako např. metoda RFWR.

Jako hlavní **teoretický přínos práce** vidím v rozboru teoretických přístupů ke kontinualizaci metody Q-učení a následné aplikaci metody lokálně vážené regrese.

Provedení srovnávací studie diskrétního a spojitého Q-učení během numerických experimentů v předloženém rozsahu, představující východisko pro jejich obecné použití, považuji za hlavní **praktický přínos práce**.

Výsledky práce byly získány v rámci projektů MSM 262100024 „Výzkum a vývoj mechatronických soustav“, pilotního projektu ÚT AV ČR č. 52020 „Řízení kráčivého robotu s využitím metod umělé inteligence“, projektu navazujícího č. 52022 „Realizace základních řídicích členů kráčivého robotu“. A za podpory výzkumného záměru CEZ: J22/98: 261100009 „Netradiční metody studia komplexních a neurčitých systémů“.

Vlastní publikace autora

[1] Březina T., Krejsa J., Věchet S.: *Stochastic Policy in Q-learning Used for Control of AMB*, pp.7-8, Inženýrská mechanika 2002, Brno

[2] Březina T., Krejsa J. Věchet S.: *Improvement of Q-learning used for controll of AMB*, Electrical Drives and Power Electronics 2003, pp.51-54, Košice, 2003

[3] Grepl R., Věchet S., Bezdíček M., Švehlák M., Chmelíček J.: *Control of experimental walking robot using simulating model*, Engineering Mechanics 2004, pp.101-102, Svratka, 2004

[4] Krejsa J., Grepl R., Věchet S.: *Approximation of walking robot stability model*, Engineering Mechanics 2004, pp. 159-160, Svratka, 2004

[5] Krejsa J., Věchet S., Pulchart J.: *Global Versus Local Approximation in Inverse Problems*, Mechatronics 2004, pp. 57-60,Warsaw

[6] Miček P., Věchet S.: *Tvorba a posuzování neuro-fuzzy modelů*, FSI Junior konference, pp.135-142, Brno, 2002

[7] Miček P., Věchet S., Březina T.: *Inverted modelling with approximation methods*, Mechatronics, Robotics and Biomachanics 2003, FSI VUT Brno, pp.123-124, 2003

[8] Miček P., Věchet S., Březina T.: *The using some approximating methods by inverse modeling*, pp.204-300, Inženýrská mechanika 2003, Brno, 2003

[9] Miček P., Věchet S., Březina T. :*Robot modeling by using quaternions*, Zborník príspevkov "Mechatronika 2003", pp.70-73, Trenčín, 2003

[10] Švehlák M., Grepl R., Věchet S., Bezdíček M., Chmelíček J.: *Design of small laboratory quadruped robot*, Engineering Mechanics 2004, Svratka, 2004

[11] Věchet S.: *Vliv parametrů modelu aktivního magnetického ložiska na algoritmus Q-učení*, Servisní robotika, pp.28, Ostrava, 2003

[12] Věchet S., Krejsa J.: *Continuous Q-learning application*, Engineering Mechanics 2004, pp. 307-308, Svratka, 2004

[13] Věchet S., Krejsa J.: *Q-learning: from discrete to continuous representation*, Mechatronics 2004, pp.12-14, Warsaw

[14] Věchet S., Krejsa J., Březina T.: *Using Modified Q-learning With LWR for Inverted Pendulum Control*,Mechatronics, Robotics and Biomechanics 2003, pp.91-92, Brno, 2003

[15] Věchet S., Miček P.: *Využití Q-učení při řízení magnetického ložiska*, FSI Junior konference, pp.235-238, Brno, 2002

[16] Věchet S., Miček P., Březina T.: *Použití modifikovaného Q-učení pro řízení inverzního kyvadla*, Engineering Mechanics 2003, pp.368-369, Prague, 2003

[17] Věchet S., Miček P., Krejsa J.: *Using Q-Learning with LWR in continuous space*, Proceedings of 6th international symposium on Mechatronics, pp.58-61, Trenčín, 2003

Použitá literatura

[18] Aha D.: *Lazy Learning*, Artificial Inteligence Review, pp. 325-337, 1997.

[19] Aljibury H.: *Improving the Preformance of Q-learning with Locally Weighted Regression.* Masters Thesis, University of Florida, 2001.

[20] An C.H., Atkeson C.G., Hollerbach J.M.: *Model-based control of a robot manipulator*, Cambrige, MA:MIT Press, 1988.

[21] Asada M., Noda S., Tawaratsumida S., Hosoda K.: *Purposive behaviour acquisition for a real robot by vision-based reinforcement learning*. Machine Learning, 279-303, 1996.

[22] Atketson C.G., Moore A.W., Schaal S.: *Locally Weighted Learning*, Artificial Inteligence Review, pp. 11-73, 1997.

[23] Atkeson C.G., Moore A.W., Schaal S.: *Locally Weighted Learning for Control*, Artificial Inteligence Review special Issue on Lazy Learning Algorithms, pp. 75-113, 1997.

[24] Atkeson C.G., Schaal S.: *Memory-based neural networks for robot learning*, Neurocomputing, vol. 9, pp. 1-27, 1995.

[25] Baird C.L.: *Residual algorithms: Reinforcement learning with functions approximation.* Machine Learning: Proceedings of the Twelfth International Conference, 1995

[26] Boone G.: *Efficient Reinforcement Learning: Model-based Acrobot Control*, International Conference on Robotics and Automation, Albuquerque, New Mexico 1997.

[27] Boyan J.A., Moore A.W.: *Generalization in Reinforcement Learning: Safely Approximating the Value Function*.

[28] Březina. T.: *Efektivní metoda Q-učení: simulační posouzení použitelnosti pro řízení aktivního magnetického ložiska*, Habilitacní a inaugurační spisy VUT v Brně, sv. 117, VUTIUM, Brno 2003 ISBN 80-214-2414-1

[29] Cook R.D.: *Influenctial observations in linear regression*. Journal of the American Statistical Association, 74, 169-174, 1979.

[30] Dahlstrom D., Wiewiora E., Cottrell G., Elkan C.: *Imitative Policies for Reinforcement Learning*.

[31] Fernandez F., Borrajo D.: *Vector Quantization Applied to Reinforcement Learning*.

[32] Forbes J., Andre D.: *Practical Reinforcement Learning in Continuous Domains*, Computer Science Division, University of California, Berkeley, Tech. Rep. UCB/CSD-001109,200.

[33] Forbes J., Andre D.: *Real-time Reinforcement Learning in Continuous Domains*, Computer Science Division, University of California.

[34] Gordon G. *Stable function approximation in dynamic programming*, International Conference on Machine Learning, 1995.

[35] Gaskett C., Wettergreen D., Zelinsky A.: *Q-learning in Continuous State and Action Spaces*.

[36] Hoder K.: Soukromé sdělení. Brno 2003

[36] Kaelbling L.P., Littman M.L., Moore A.W.: *Reinforcement Learning: A Survey*.1995

[37] Kretchmar R.M., Young P.M., Anderson C.W., Hittle D.C., Anderson M.L., Delnero C.C.: *Robust Reinforcement Learning Control with Static and Dinamic Stability*, Colorado State University, 2000.

[38] Lewandowski A., Tagscherer M., Kindermann L., Protzel P.: *Improving the Fit of Locally Weighted Regression Models*,

[39] Lin L.J.: *self-improving reactive agents based on reinforcement learning, planning and teaching*. Machine Learning, 8, 293-321, 1992.

[40] Marada T.: *Využití Q-učení pro řízení asynchronního elektromotoru*, Pojednání ke státní doktorské zkoušce, Brno 2003.

[41] Moore A.W.: *Efficient memory-based learning for robot control*, Computer Laboratory University of Cambridge, 1990.

[42] Munos R.: *A convergent Reinforcement Learning algorithm in the continuous case: the Finite-Element Reinforcement Learning*, International Conference on Machine Learning 1996 (ICML '96), 1996

[43] Munos R.: *A convergent Reinforcement Learning algorithm in the continuous case based on a Finite Difference method*, International Joint Conference on Artificial Intelligence (IJCAI '97), 1997

[44] Nakanishi J., Farrell J.A., Schaal S.: *A Locally Weighted Learning Composite Adaptive Controller with Structure Adaptation*,

[45] Nikovsky D.: *Visual Memory-based Learning for Mobile Robot Navigation*, Second Conference on Computational Inteligence and Neurosciences, vol. 2, pp.1-4, 1997.

[46] Půst, L.: *Dynamická stabilita magnetického ložiska.* Engineering Mechamics'97, Svratka 1997, pp.57-62.

[47] Schaal S., Atkeson C.G.: *Constructive Incremental Learning From Only Local Information*, Neural Computation, 10, 8, 2047-2084.

[48] Schaal S., Atkeson C.G.: *Robot juggling: An implementation of memory-based learning*, Control Systems Magazine, vol. 14, pp. 57-71, 1994.

[49] Schaal S., Atkeson C.G., Vijayakumar S.: *Real-time Robot Learning With Locally Weighted Statistical Learning*, International Conference on Robotics and Automation, San Francisko, 2000.

[50] Schaal S., Atkeson C.G., Vijayakumar S.: *Scalable Techniques from Nonparametric Statistics for Real Time Robot Learning*.

[51] Singh S.P., Sutton R.S.: *Reinforcement learning with replacing eligibility traces*. Machine Leraning, 22, 123-158, 1996.

[52] Smart W.D., Kaelbling L.P.: *Practical Reinforcement Learning in Continuous Spaces*.

[53] Smart W.D., Kaelbling L.P.: *Reinforcement Learning for Robot Control*.

[54] Sutton R.S.: *Generalization in reinforcement learning: successful examples using sparse coarse coding*. Advances in Neural Information Processing systems, pp. 1038-1044, 1996.

[55] Szepesvári C.: *Convergent Reinforcement Learning with Value Function Interpolation*.

[56] Szepesvári C., Littman M.L.: *A Unified Analyis of Value-function-based Reinforcement-learning Algorithms*. Neural Computation, pp. 2017-2059, 1999.

[57] Tadepalli P., Ok D.: *Scaling up Average Reward Reinforcement Learning bz Approximating the Domain Models and the Value Function*,13th Int. Conf. On Machine Learning, pp. 471-479, 1996.

[58] Takahashi Y., Takeda M., Asada M.: *Improvement Continuous Valued Q-learning and its Application to Vision Guided Behavior Acquisition*.

[59] Thrun S., Schwartz A.: *Issues in using function approximation for reinforcement learning*.Proceedings of the Fourth Connectionist Models Summer School, 1993.

[60] Vijayakumar S., Schaal S.: *Locally Weighted Projection Regression: An O(n) Algorithm for Incremental Real Time Learning in High Dimensional Space*.

[61] Vijayakumar S., Schaal S.: *Real Time Learning in Humanoids: A Chalange for Scalability of Online Algorithms*.

[62] Watkins C.J., Dayan P.: *Q-learning*, Machine learning, 8, 279-292, 1992.

[63] Watkins, C., J., C., H.: *Learning from delayed rewards*. Ph.D. Thesis, Cambridge University, Cambridge, England, 1989.

[64] Williams R.J.: *Simple statistical gradient-following algorithms for connestionist reinforcement learning*, Machine Learning, 8, 229-256, 1992.

Curriculum Vitae

Ing. Stanislav Věchet Tel: 541 142 874
Bosonožská 17, Email: vechet.s@fme.vutbr.cz
Brno, 625 00

Datum narození: 22. duben 1977
Národnost: Česká
Občanství: ČR

Vzdělání

od roku 1990 Střední průmyslová škola strojnická, Sokolská 1, Brno,
ukončeno maturitou 1995

od roku 1995 VUT FSI v Brně, Ústav mechaniky těles, Technická 2, Brno
magisterské studium, obor Mechatronika,
ukončeno státní závěrečnou zkouškou 2000

od roku 2000 VUT FSI v Brně, Ústav automatizace a informatiky,
obor Informatika, plánované ukončení jaro 2005

od roku 2001 VUT FSI v Brně, Ústav mechaniky těles
obor Mechatronika, doktorské studium,
plánované ukončení podzim 2004

Praxe

od roku 2000 Programování driverů a testování mikroprocesorů
Motorola, UNIS, spol. s r.o., reference: Ing. Martin Stročka

od roku 2002 Programátorské práce, konzultační činnost, JmSKe,
reference: Ing. Petr Vrbík

Jazykové znalosti

Angličtina: aktivní znalost, Němčina: pasivní znalost, Ruština: základy

Summary

Presented PhD thesis deals with extension of classical Q-learning algorithm, which works with discrete sets of states and actions, into continuous space. The thesis discusses the continualization of Q-learning using local approximators, locally weighted regression method (LWR) specifically. The LWR algorithm was primarily implemented and tested independently on various test data. After those tests it was used for continualization of Q-learning algorithm. To verify the properties of continuous Q-learning two simulation models were prepared: inverse pendulum and active magnetic bearing models. Those models were used both in initial stages focused on first attempts with continuous Q-learning and in stages of simulation verification of the method itself.

During the initial stages of testing the standard simulation conditions were found and consequently the main simulation stage was performed, focused on identification of influences of particular parameters on given discrete and continuous Q-learning algorithm.

The main contribution of continuous Q-learning is the elimination of the problems with discretization of state space. Due to the possibility of using the approximator parameters setting independently on used model the use of the method for various simulation models is greatly simplified. Continuous Q-learning also reached higher learning speed while keeping the control quality criterion values , which lead to reduction of the time necessary for testing the properties of given model.

Final words refer to possible further development in verification simulations and replacement of used method of locally weighted regression by more advanced method.