Národní konference s mezinárodní účastí

INŽENÝRSKÁ MECHANIKA 2002

13. – 16. 5. 2002, Svratka, Česká republika

STOCHASTIC POLICY IN Q-LEARNING USED FOR CONTROL OF AMB

Tomáš Březina[1], Jiří Krejsa[2], Stanislav Věchet[3]

Abstract

A great intention is lately focused on Reinforcement Learning (RL) methods. The article is focused on improving model free RL method known as Q-learning algorithm used on active magnetic bearing (AMB) model. Stochastic strategy and adaptive integration step increased the speed of learning approximately hundred times. Impossibility of using proposed improvement online is the only drawback, however it might be used for pretraining on simulation model and further fined online.

Keywords: *Reinforcement Learning, Q-learning, Active Magnetic Bearing*

1. Úvod

V poslední době je značná pozornost věnována metodě opakovaně posilovaného učení (reinforcement learning, RL). Jedním z nejpopulárnějších a nejefektivnějších RL algoritmů bez modelu je Q-učení. Této koncepce bylo použito například na simulační úlohu řízení jednohmotového modelu aktivního magnetického ložiska (AML), viz. [1]. Následující článek studuje problematiku vlivu stochastické strategie použité v Q-učení na této úloze.

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 okamžitou funkci odměny $r(s_t, s_{t+1}, a_t)$, která určuje konkrétní odměnu/pokutu za přechod ze stavu s_t do stavu s_{t+1}.

Dlouhodobý cíl agenta je definován jako funkce okamžitých odměn, například kumulativní srážková odměna (cumulative discount reward)

[1] Tomáš Březina, Ing. RNDr. CSc.,VUT v Brně, FSI ÚAI, Technická 2, 616 69 Brno, brezina@uai.fme.vutbr.cz
[2] Jiří Krejsa, Ing. PhD., ÚT AV ČR, Technická 2. 616 69, Brno
[3] Stanislav Věchet, Ing., VUT v Brně, FSI, ÚMT, Technická 2, 616 69, Brno

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

kde $0 \le g < 1$ 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 $a = p(s)$ a cílem opakovaně posilovaného učení je najít optimální strategii p^*, která maximalizuje kumulativní srážkovou odměnu. Pro účel nalezení optimální strategie zavádíme hodnotovou funkci $f^p(s)$ strategie p, která udává očekávanou kumulativní srážkovou odměnu při počátečním stavu s a použití strategie p.

3. Q-učení

Při Q-učení je hodnotová funkce $f^p(s)$ 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:

$$Q(s_t, a_t) = (1-a)Q(s_t, a_t) + a\left(r(s_t, a_t, s_{t+1}) + g \max_{a_t} Q(s_{t+1}, a_t)\right)$$

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. Aplikace

Jako modelu prostředí je použito jednoduchého jednohmotového modelu AML [2]

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

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 $F_m(x(t), i(t))$ je síla od magnetického ložiska působící na hmotný bod, aproximovaná

$$F_m(x, I; \bar{c}) = -c_1 x + c_2 I - c_3 x^3 + c_4 I x^2 - c_5 I^2 x - c_6 I^3, \quad c_1, \mathbf{K}, c_6 > 0.$$

Úkolem agenta je, aby svým 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.

Pro popis prostředí byl zvolen třírozměrný stavový prostor s proměnnými $x, \dot{x}, \ddot{x}$. Jako akce agenta bylo zvoleno napájecí napětí u (akce byly brány z množiny $\{-U_{\max}, 0, U_{\max}\}$). Q-funkce byla aproximována tabulkou (čtyřrozměrným polem) s použitím nelineárního rastru.

Počáteční podmínky x_0, $\dot{x}_0$, $\ddot{x}_0$ byly voleny náhodně, ale pouze takové, pro které je soustava "řiditelná", tj. použití napětí $U_{\max}$ (s vhodným znaménkem) zaručuje existenci takového času t, pro který $x(t) = 0$.

5. Stochastická strategie

Běžnou strategií při použití Q-učení je využívat k učení všech stavů prostředí, které nastanou v průběhu jednotlivých pokusů postupně. V takovém případě ovšem může dojít k tomu, že se v průběhu učení vytvoří pásma, ve kterých se bude systém pohybovat nejčastěji a ostatní uzlová místa (akce v určitých stavech) budou použita méně často, případně vůbec.

Vzhledem k výše uvedenému byla použita pro procházení jednotlivých uzlových bodů stochastická strategie. V každém kroku učení jsou použity všechny kombinace stavů a akcí (ve všech stavech jsou provedeny všechny možné akce) právě jednou s využitím pouze jednoho přechodu jednotlivých stavů. Jednotlivé kombinace stavů jsou předkládány při učení jak sekvenčně postupným průchodem všech uzlů rastru tabulky podle jednotlivých rozměrů, tak v náhodném pořadí.

Při učení je žádoucí, aby během jednoho kroku systém přešel ze stavu reprezentovaného jedním uzlovým místem v tabulce aproximující Q-funkci do stavu jiného. Proto byl při učení použit adaptivní integrační krok: Integrační krok je opakovaně prodlužován, dokud nedojde ke změně stavu nebo integrační krok nepřesáhne určitý limit.

6. Provedené experimenty

Pro experimenty bylo použito hodnot parametrů modelu podle [2] ($m = 5,65\,[\mathrm{kg}]$, $R = 2\,[\Omega]$, $L = 5 \cdot 10^{-3}\,[\mathrm{H}]$, $b = 20\,[\mathrm{kgs^{-1}}]$, $c_1 = 1289 \cdot 10^3\,[\mathrm{kgs^{-2}}]$, $c_2 = 454\,[\mathrm{kgms^{-2}A^{-1}}]$, $c_3 = 3184 \cdot 10^9\,[\mathrm{kgm^{-2}s^{-2}}]$, $c_4 = 841 \cdot 10^6\,[\mathrm{kgm^{-1}A^{-1}s^{-2}}]$, $c_5 = 38 \cdot 10^3\,[\mathrm{kgs^{-2}A^{-2}}]$, $c_6 = 30\,[\mathrm{kgms^{-2}A^{-3}}]$). Hodnoty x_r a $\dot{x}_r$ byly zvoleny $x_r = 10.10^{-6}\,[\mathrm{m}]$, $\dot{x}_r = 20.10^{-3}\,[\mathrm{ms^{-1}}]$, $U_{\max} = 100[\mathrm{V}]$.

V experimentech byl sledován vliv:

1. Stochastické strategie: účelem bylo zjistit, zda je tato strategie vůbec použitelná a jakým způsobem ovlivňuje rychlost učení.
2. Adaptačního integračního kroku: účelem bylo zjistit, zda použití adaptačního integračního kroku (a tím zvýšení počtu přechodů ze jednoho uzlového bodu do dalšího) zlepší kvalitu řízení.
3. Tvaru posilovací funkce: Posilovací funkce byla použita ve dvou variantách. První stanovuje odměnu/pokutu pouze z množiny $\{-1,0,1\}$, druhá používá pro odměnu lineárního vztahu v závislosti na absolutní hodnotě x.

4. Velikosti napájecího napětí (akční veličiny): Původní maximální hodnota U_{max} byla postupně snižována na 60 a 30 voltů.

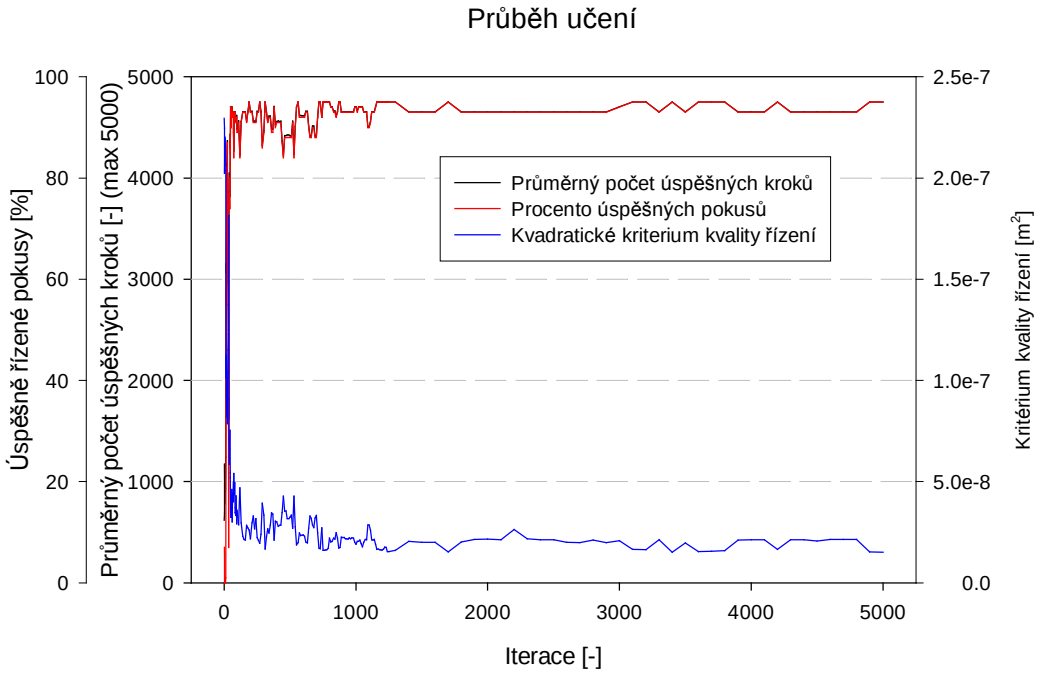
6.1 Hodnocení testů

Systém byl testován v průběhu učení následujícím způsobem: Nejprve bylo vygenerováno náhodně *n* počátečních stavů splňujících podmínku regulovatelnosti. Tyto stavy byly použity během testování (s náhodným zatížením) a bylo sledováno, zda po stanovený počet kroků dokáže agent udržet výchylku hmotného bodu v určeném intervalu. Pokud to agent dokázal, byl pokus vyhodnocen jako úspěšný. Tím byla stanovena procentuální úspěšnost v průběhu učení. Dále byla sledována hodnota kvadratického kriteria kvality regulace a to jak celková tak i odděleně pouze pro úspěšné pokusy.

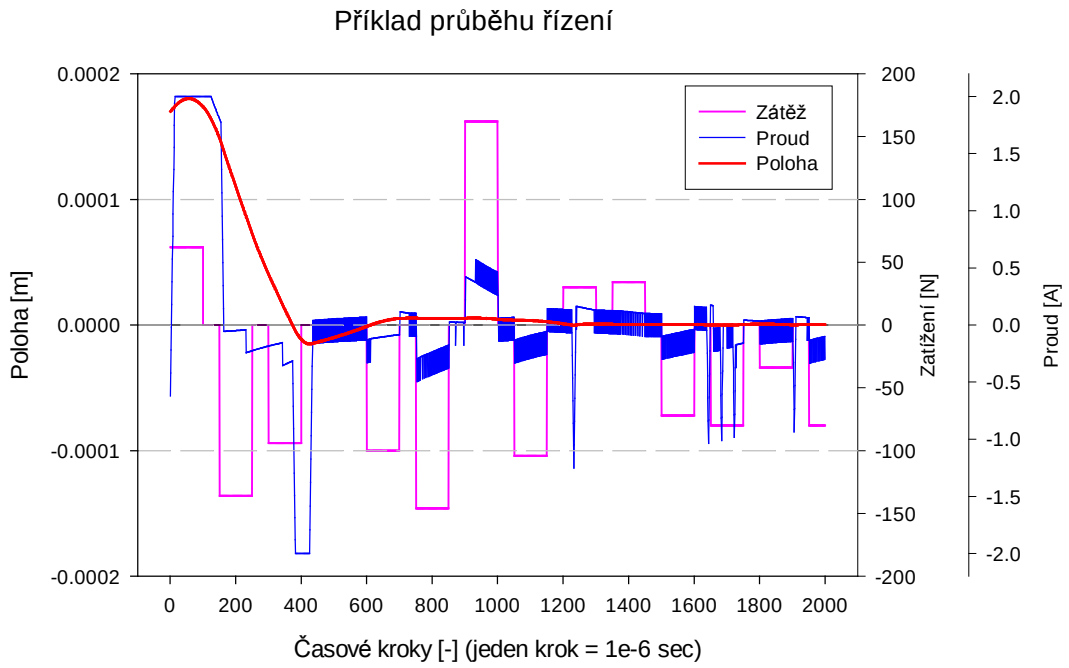
6.2 Průběh učení

Zhodnocení stochastické strategie

Pro zhodnocení stochastické strategie bylo použito následující nastavení: velikost rastru tabulky aproximující Q-funkci: 12x12x10 (poloha, rychlost, zrychlení), adaptivní integrační krok, sekvenční průchod tabulkou, lineární formulace posilovací funkce. Toto nastavení je dále považováno za standard. Průběh učení je zobrazen na Obr. 1. Z obrázku je zřejmé že k naučení dojde během přibližně 100 iterací. Další experimenty ověřující vliv ostatních parametrů byly proto ukončeny již po 300 iteracích. Příklad řízení při konkrétním průběhu zátěžné síly a konkrétních počátečních podmínkách je uveden na Obr. 2. Na obrázku je zobrazena poloha, průběh zátěžné síly a průběh proudu.



Obr. 1. Průběh učení při standardním nastavení Q-učení



Obr. 2. Průběh polohy, zátěže a proudu během řízení

Vliv adaptačního integračního kroku

Pokud je použita fixní hodnota integračního kroku 1e-6, dojde během jednoho kompletního průchodu tabulkou k 5 až 10 změnám stavů. V takovém případě řídící člen dosahuje zhruba 35% úspěšnost ze všech testovacích pokusů a nastavení ostatních parametrů Q-učení tento výsledek již nezlepší. Při použití adaptačního integračního kroku dojde k nárůstu počtu změn stavů na 250-280 a úspěšnost dosahuje 95-100 %.

Vliv tvaru posilovací funkce

Posilovací funkce agenta r byla definována v základním tvaru

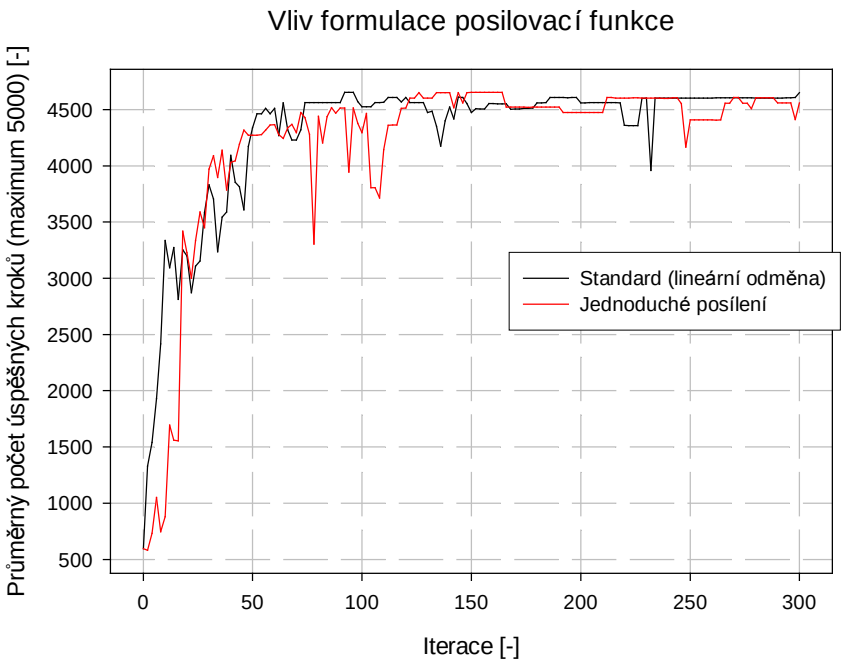
$$r = \begin{cases} 1 & \text{pro } \left(|x_k| \le x_r\right) \cap \left(|\dot{x}_k| \le \dot{x}_r\right) \\ 0 & \text{pro } \left(\left(x_r < |x_k| < x_{\max}\right) \cup \left(\dot{x}_r < |\dot{x}_k|\right)\right) \\ -1 & \text{jinak} \end{cases}$$

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

Druhá varianta posilovací funkce počítá s využitím lineárního vztahu pro odměnu

$$r = \begin{cases} 1 - \left(|x_k| / x_r\right) & \text{pro } \left(|x_k| \le x_r\right) \cap \left(|\dot{x}_k| \le \dot{x}_r\right) \\ 0 & \text{pro } \left(\left(x_r < |x_k| < x_{\max}\right) \cup \left(\dot{x}_r < |\dot{x}_k|\right)\right) \\ -1 & \text{jinak} \end{cases}$$

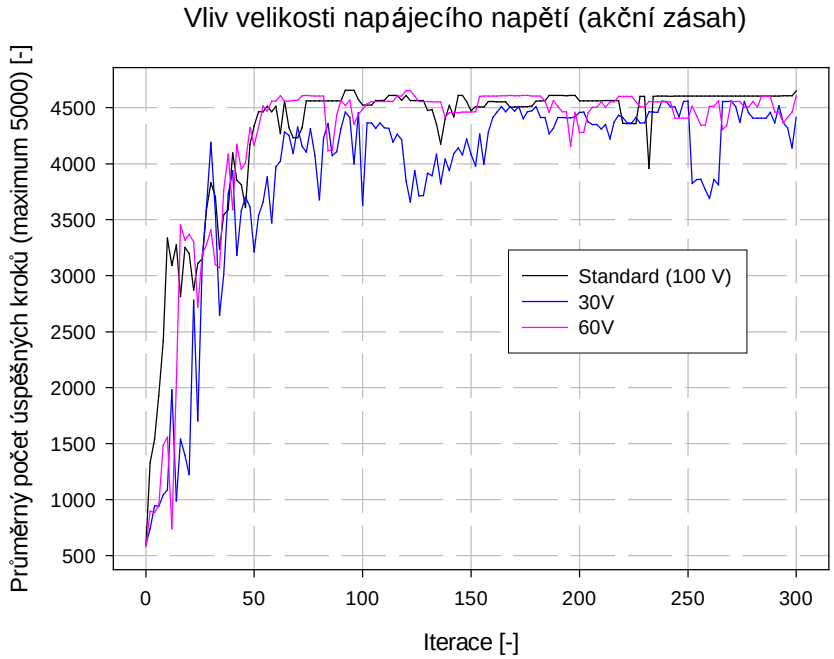
Srovnání průběhu učení při použití obou variant tvaru posilovací funkce je zobrazeno na Obr. 3. Je zřejmé že vliv tvaru posilovací funkce není velký, nicméně varianta s lineární odměnou vykazuje stabilnější průběh.



Obr. 3. Vliv tvaru posilovací funkce

Vliv velikosti napájecího napětí (akční veličiny)

Maximální hodnota napájecího napětí původně stanovená na 100 V byla postupně snižována na 60 a 30 V. Průběh učení s těmito hodnotami je zobrazen na Obr. 4. Z grafu je zřejmé že snížení na 60 V nevnáší do průběhu učení velkou změnu, nicméně použití 30 V snižuje v souladu s předpoklady rychlost a kvalitu učení.



Obr. 4. Vliv velikosti napájecího napětí

7. Závěr

Vzhledem ke zcela rozdílnému konceptu není možné přesně porovnat stochastickou strategii procházení uzlových bodů se strategií použitou v minulých experimentech, nicméně lze říci že použití stochastické strategie a adaptivního integračního kroku se projevilo ve výrazném snížení časové náročnosti učení (přibližně stokrát).

Při použití stochastické strategie se Q-učení projevuje jako robustní metoda, která prokazuje nízkou citlivost i na nastavení vlastních parametrů metody (použití permutací, nastavení srážkového faktoru, definice opakovací funkce).

Nevýhodou použité strategie je nemožnost použití online učení, je ji ovšem možné použít pro předtrénování na simulačním modelu a takto připravený řídící člen již dále doučovat online.

Poděkování

Tato práce je podporována projektem GAČR č. 101/00/1471 a výzkumným záměrem MŠMT MSM 262100024

Použitá literatura

[1] Březina T., Ehrenberger Z., Kratochvíl C.. Reinforcement Learning Model: Control of Nonlinear and Unstable Processes, Engineering Mechanics 2001, Svratka, ČR, 2001

[2] Půst L.: Dynamic stability of magnetic bearing, Engineering Mechanics 97, Svratka, pp. 57-62, 1997