ENGINEERING MECHANICS 2004
NATIONAL CONFERENCE
with international participation
Svratka, Czech Republic, 10 - 13 May 2004

# USE OF CONTINUOUS ACTION REINFORCEMENT LEARNING AUTOMATA FOR ASYNCHRONOUS ELECTROMOTOR CONTROL

T. Březina*, M. Turek*

**Summary:** *Relatively unknown reinforcement learning algorithm, so called continuous action reinforcement learning automaton, is presented in this contribution. Automaton learning algorithm is based on rewarding, that gradually evolves set of probability densities. This set is consequently used for action set determination. Simulation study describing learning and behavior of asynchronous electromotor control is further presented. Standard PSD controller is used whose parameter values represent actions of three independent automata. The aim of online learning process is to minimize mean square of control error. Here described learning algorithm is simple to implement, robust to high level of noise.*

## 1. Úvod

Přes veškerý vývoj v oblasti řídicích systémů patří PID regulátor stále mezi nejpoužívanější. Obliba PID regulátoru je způsobena jeho všestranností, vysokou spolehlivostí a jednoduchostí.

Standardní tvar PID regulátoru definuje akční veličinu $u(t)$ váženým součtem tří dynamických funkcí regulační odchylky $e(t)$

$$e(t) = y(t) - y_{ref}(t),$$
$$u(t) = K_p \cdot e(t) + K_i \cdot \int_0^t e(t)dt + K_d \cdot \frac{de(t)}{dt}, \tag{1}$$

kde $y_{ref}(t)$ je referenční (požadovaný) průběh výstupu soustavy a $y(t)$ je měřený výstup soustavy zatížený jednak chybami měření (šumem senzorů) a jednak vlivem poruch soustavy, o kterých se předpokládá, že jsou neznámé. Parametry PID regulátoru $K_p$, $K_i$ a $K_d$ jsou nastaveny tak, aby regulovaná soustava splňovala požadovaná kritéria chování, obvykle vymezená velikostí překmitu, dobou ustálení a velikostí regulační odchylky v ustáleném stavu při skokové změně referenční veličiny $y_{ref}$.

Pro nastavení hodnot parametrů regulátoru existuje celá řada metodik. Z nich jsou zvláště atraktivní metodiky, které nepotřebují model regulované soustavy, ale pro nastavování používají přímo soustavu. Standardní a stále oblíbenou metodikou, která nepotřebuje model soustavy, je metodika Ziegler-Nicholsova. Na jednoduchých soustavách dává dobré výsledky. Naopak u složitých soustav mohou být takto získané hodnoty parametrů PID daleko od optimálních.

* doc. Ing. RNDr. Tomáš Březina, CSc., Milan Turek, VUT v Brně, FSI, ÚAI, Technická 2, 616 69 Brno, brezina@uai.fme.vutbr.cz

Jinou možností je ladění parametrů s využitím metod, které minimalizují určité kritérium technické charakteristiky řízené soustavy. Pro minimalizaci kritéria, typicky formulovaného jako střední kvadratická odchylka mezi dosaženým a požadovaným průběhem jisté veličiny (veličin) soustavy, se často používá gradientních metod. Jestliže kritérium vykazuje více minim, pak gradientní metody používající střední kvadratickou odchylku nemusí konvergovat ke globálnímu minimu. Naproti tomu metody stochastické optimalizace, např. genetický algoritmus (Holland, 1975), simulované žíhání nebo učící se automaty (Narendra & Thathachar, 1989) zvyšují pravděpodobnost nalezení globálního minima. Problémem však může být pomalá konvergence těchto metod.

V tomto příspěvku je popsána relativně málo známá jednoduchá metoda (CARLA). Metoda CARLA vychází z formalismu stochastických diskrétních učících se automatů a tento formalismus rozšiřuje na spojitou oblast. Byla již úspěšně použita pro nastavení řídicího členu volnoběhu spalovacího motoru (Howell & Best, 2000), odpružení závěsu kola osobního automobilu (Howell, Frost, Gordon & Wu, 1997) a návrhu adaptivního digitálního filtru (Howell & Gordon, 2001). V příspěvku jsou dále uvedeny výsledky jejího simulačního ověření při ladění parametrů PSD regulátoru pro řízení otáček malého asynchronního elektrického motoru. Metoda nevyžaduje model řízené soustavy. Díky své jednoduchosti a dobrým konvergenčním vlastnostem je použitelná jako adaptivní systém, který se učí v reálném čase.

## 2. Diskrétní stochastické učící se automaty

Stochastické učící se automaty (Narendra & Thathachar, 1989; Najim & Poznak, 1994) jsou jedním z reprezentantů opakovaně posilovaného učení. Mohou pracovat v náhodném a neznámém prostředí. Náhodně vybírají akce z konečné diskrétní množiny akcí $\{x_1, x_2, \mathbf{K}, x_r\} \in X$ podle interního ohodnocení pravděpodobnostmi $p_i$ jejich použití v prostředí. Z prostředí je následně jako odezva přijat posilovací signál $b$, který ohodnocuje úspěšnost právě použité akce z hlediska cíle provádění zásahů do prostředí. Posilovacího signálu je použito k aktualizaci interního rozložení pravděpodobnosti automatu tak, aby podle konkrétního pravidla učení byly úspěšným akcím zvyšovány pravděpodobnosti jejich použití, zatímco ostatním akcím byly pravděpodobnosti ponechány beze změny nebo snižovány. Akce s nejvyšší pravděpodobností nakonec odpovídá při vhodné volbě $b$ globálnímu minimu (důkaz konvergence viz. Narendra & Thathachar, 1989; Najim & Poznak, 1994). Jednou z možností inženýrského použití je ztotožnit množinu akcí $X$ s množinu možných hodnot parametrů, např. řídicího členu, při vhodné volbě posilovacího signálu $b$, který bude reflektovat např. hodnotu kritéria kvality řízení. Při této organizaci se až do určité meze bude postupně zlepšovat průměrné chování řízené soustavy. Učící se automat je při tomto použití možno chápat jako „chytrý“ parametr, který je schopen sám ladit svoji hodnotu.

Existuje mnoho pravidel učení s různými vlastnostmi konvergence (Narendra & Thathachar, 1989). Jedním z nejpoužívanějších algoritmů je schéma lineární odměny/nečinnosti $L_{RI}$, o kterém bylo dokázáno, že má vyhovující konvergenční vlastnosti (Narendra & Thathachar, 1989). Jako odezva na akci $x_i$, která byla vybrána v časovém kroku $n$, jsou pravděpodobnosti použití akcí aktualizovány podle

$$p_i(n+1) = p_i(n) + qb(n)\left(1 - p_i(n)\right)$$
$$p_i(n+1) = p_i(n) + qb(n)\, p_j(n), \qquad \text{pro } i \neq j, \tag{2}$$

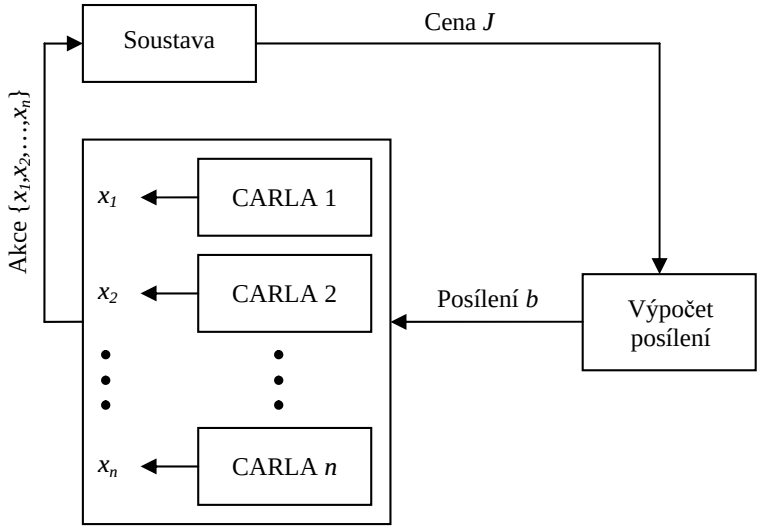
kde je $q$, $0 < q < 1$ parametr učení a $b \in [0,1]$ je posilovací signál vyjadřující hodnotu odměny přijaté pro akci (s krajními hodnotami - 1 pro maximální odměnu a 0 pro prázdnou (žádnou) odměnu).

Jediný učící se automat je relativně jednoduchou jednotkou. Může však být použit i ve složitějších systémech s vícesložkovými akcemi. Pro proměnnou spojité akce se získá odpovídající množina diskrétních akcí diskretizací původního intervalu spojité akce. Rychlost konvergence je však tím nižší, čím vyšší je počet akcí (s velmi nízkými počátečními pravděpodobnostmi). To může být řešeno paralelním propojením učící se automatů, kdy každý z automatů vybírá pouze s malého počtu akcí. Toto schéma je rovněž použitelné i v případě vícesložkových akcí.

Podstatou metody CARLA je rozšíření naznačeného diskrétního přístupu, které vede na spojitý tvar učícího se automatu.

## 3. Metoda CARLA

Dále uvedená formulace CARLA algoritmu (Continuous Action Reinforcement Learning Automaton) uvažuje případ, kdy na prostředí působí jediná akce. Poznamenejme, že vícesložkové akce jsou typicky konfigurovány podle obr. 1.




Obr. 1: Typická konfigurace pro vícesložkové akce

Nechť proměnná $x$ akce algoritmu CARLA je omezená spojitá náhodná proměnná definovaná na intervalu $X = [x_{min}, x_{max}] \subset \mathbf{R}$. Diskrétní rozložení pravděpodobností stochastického učícího se automatu je v $n$-té iteraci, $n = 1, 2, .\mathbf{K}$, nahrazeno spojitým rozložením pravděpodobnosti $f(x, n) \geq 0$, $\forall n$, $\forall x$, pro které

$$\int_{-\infty}^{\infty} f(x, n)\, dx = \int_{X} f(x, n)\, dx = 1. \tag{3}$$

Počáteční rozložení je voleno jako rovnoměrné

$$f(x,0)=\begin{cases}\dfrac{1}{x_{max}-x_{min}}, & pro\ x\in X,\\ 0, & jinak.\end{cases} \tag{4}$$

Hodnota akce $x(n)$ je vybírána použitím rozložení $f(x,n)$. Toho se dosáhne prostřednictvím pomocné náhodné proměnné $z(n)$ s rovnoměrným rozložením $U[0,1]$. Hodnota akce $x(n)$ se při hodnotě $z(n)$ určí podle

$$\int_{x_{min}}^{x(n)} f(x,n)\,dx = z(n). \tag{5}$$

Akce $x(n)$ je potom použita v prostředí, jehož odezvou je cena (ohodnocení funkční charakteristiky soustavy) $J(n)$, která je mírou kvalitu použité akce. Posílení $b(n)\in[0,1]$, se vypočte jako

$$b(n)=\max\left\{0,\frac{J_{stř}-J(n)}{J_{stř}-J_{\min}}\right\}, \tag{6}$$

kde $J_{stř}$ a $J_{min}$ jsou střední a minimální hodnota cen z referenční množiny. Posílení $b(n)$ tak vyjadřuje cenu $J(n)$ poslední akce s ohledem na ceny předchozích akcí. Vysoká hodnota $b(n)$ značí odměnu a nízká hodnota nečinnost. Posledních $m$ hodnot ceny akcí je evidováno v referenční množině $R$. Omezení evidence na posledních $m$ cen akcí není dáno pouze implementačním omezeními. Velikost této hodnoty ovlivňuje rychlost adaptace systému na měnící se prostření.

Rozložení pravděpodobnosti je aktualizováno podle pravidla učení

$$f(x,n)=\begin{cases}a\left[f(x,n)+b(n)H(x,r)\right], & pro\ x\in X\\ 0, & jinak,\end{cases} \tag{7}$$

přičemž $a$ je určeno normalizační podmínkou

$$\int_{x_{min}}^{x_{max}} f(x,n)\,dx = 1. \tag{8}$$

V (7) značí $H(x,r)$ (symetrickou) Gaussovu funkci se středem v $r=x(n)$

$$H(x,r)=l\,\exp\left(-\frac{(x-r)^2}{2s^2}\right) \tag{9}$$

a $l>0$ a $s>0$ jsou parametry, které ovlivňují výšku a šířku Gaussovy funkce. Jsou definovány vzhledem k intervalu akce, např. jako

$$l = \frac{g_h}{x_{max} - x_{min}}, \qquad s = g_w \left( x_{max} - x_{min} \right), \tag{10}$$

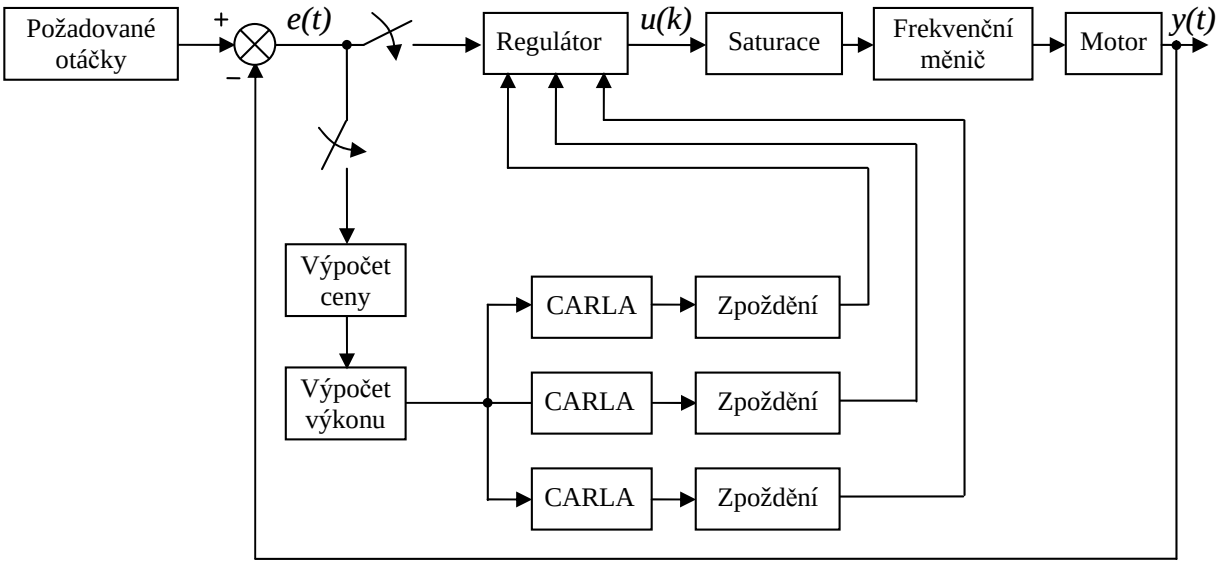
kde je rychlost resp. rozlišení učení řízeno dvěma nezávislými parametry $g_h > 0$ resp. $g_w > 0$.

## 4. Simulační ověření

K simulacím bylo použito úplného dynamického modelu malého asynchronního motoru podle (Ong, 1998) s hodnotami parametrů podle tab. 1.

| $R_r\,[\Omega]$ | $R_s\,[\Omega]$ | $L_r\,[H]$ | $L_s\,[H]$ | $L_{rs}\,[H]$ | $J\left[kg.m^2\right]$ |
|---|---|---|---|---|---|
| 2.95 | 4.37 | 0.476 | 0.471 | 0.459 | 0.0015 |

Tab. 1: Parametry asynchronního elektromotoru 4AP 90S-2




Obr. 2: Schéma regulované soustavy

Schéma celé řízené soustavy je uvedeno na obr. 2. Pro řízení motoru je použito diskrétního proporcionálně-sumačně-derivačního (PSD) regulátoru. Výpočet akční veličiny regulátoru je prováděn podle

$$u(n) = K_p \cdot e(n) + K_s \cdot T \cdot \sum_{i=1}^{n} e(i) + \frac{K_d}{T} \cdot \left( e(n) - e(n-1) \right), \tag{11}$$

kde $T$ je perioda vzorkovače a $K_p$, $K_s$ a $K_d$ jsou parametry (proporcionální, sumační a derivační) PSD regulátoru. Každý ze tří automatů CARLA nastavuje velikost jednoho z parametrů regulátoru. Za regulátorem je zařazena saturace jako omezovač požadované hodnoty frekvence měniče na rozsah 0 až 50 Hz. Frekvenční měnič zároveň definuje rychlost změny frekvence (rampa). Zpoždění za automaty CARLA je zařazeno pro zohlednění doby kroku výpočtu algoritmu CARLA (dopravního zpoždění mezi vstupem regulační odchylky do řídicího členu a výstupem nových hodnot parametrů PSD regulátoru).

Jako ceny akce je použito obvyklého kritéria

$$J(n) = \left[e(n)\right]^2 . \tag{12}$$

| | | |
|---|---|---|
| Časový krok simulace | $10^{-5}$ | [s] |
| Chyba snímače otáček | $\pm 2$ | [$\text{min}^{-1}$] |
| Perioda vzorkovače regulátoru | $10^{-4}$ | [s] |
| Přesnost vzorkovače regulátoru | $\pm 1$ | [$\text{min}^{-1}$] |
| Perioda vzorkovače CARLY | $10^{-2}$ | [s] |
| Přesnost vzorkovače CARLY | $\pm 10^{-1}$ | [-] |
| Zpoždění | $10^{-3}$ | [s] |
| Rampa zdroje požadovaných otáček | 105 | [$\text{s}^{-1}$] |
| Rampa frekvenčního měniče | 650 | [$\text{Hz.s}^{-1}$] |
| Interval akcí $K_p$ | $< 4.4, 8.3 > 10^{-2}$ | [Hz.s] |
| Interval akcí $K_s$ | $< 5.0; 6.1 > 10^{-1}$ | [Hz] |
| Interval akcí $K_d$ | $< 1.2; 2.3 > 10^{-3}$ | [$\text{Hz.s}^2$] |

Tab. 2: Parametry modelu řízené soustavy

## 5. Chování metody

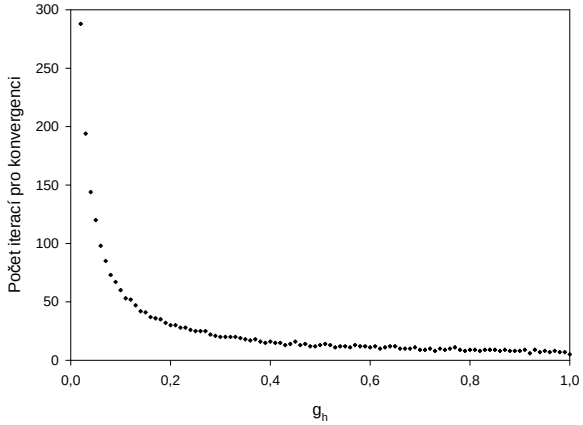
Vyhledávacími simulacemi bylo zjištěno, že parametry $g_h$ a. $g_w$. rovnic (10), které ovlivňují aktualizaci rozložení pravděpodobnosti úspěšnou akcí, významně ovlivňují proces konvergence.

Hledání vhodné hodnoty parametru $g_h$ bylo prováděno s jediným automatem CARLA, jehož úkolem bylo naladit se na konstantní hodnotu $0,2$ (eventuálně se přeladit při skokové změně z hodnoty $0,2$ na hodnotu $0,8$). Pro konkrétní hodnoty $g_h$ byla zjišťována první iterace $n$, pro kterou platí

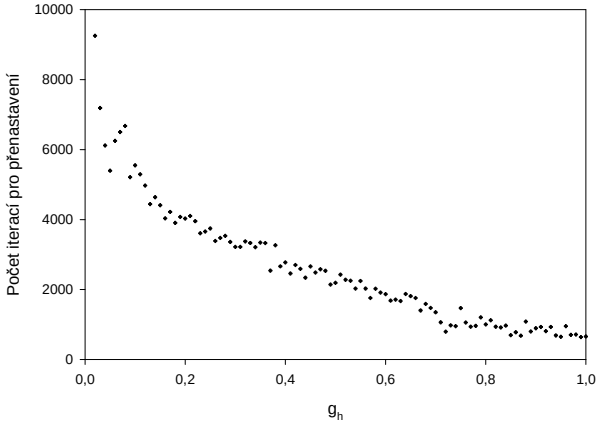
$$\Delta x(n) \le 0,2 , \tag{13}$$

kde $\Delta x(n) = \max X' - \min X'$ a $X' = \left\{x; x \in [0,1] \text{ a } f(x,n) \ge 0,1\right\}$.

Čím vyšší je hodnota parametru $g_h$, tím bližší jsou následně vybírané akce poslední úspěšné akci a zrychluje se konvergence (viz. obr. 3a). Přitom je vhodné, aby plně úspěšná akce ($b = 1$) sice měla značný vliv na rozložení pravděpodobnosti, ale pouze takový, aby nepotlačila účinky celého předcházejícího procesu učení. V takovém případě se zvýší ochota metody přelaďovat hodnoty parametrů PSD regulátoru podle okamžitých posílení (a tím se sníží stabilita celé metody), viz. obr. 3b. Jako kompromisní se nejlépe osvědčila volba $g_h = 0.2$

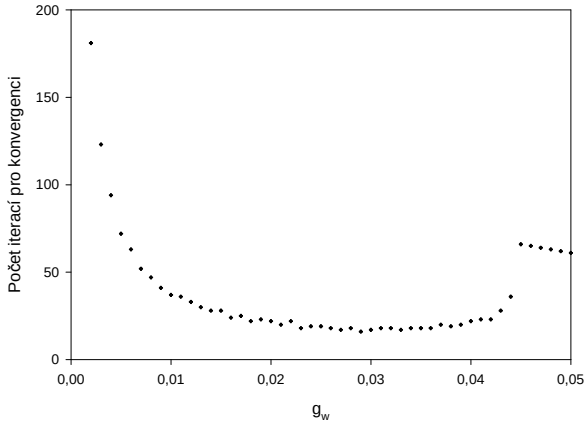



Obr. 3a: Vliv parametru $g_h$ na počet iterací nutný k naladění konst hodnoty $0,2$ za podmínek podle (13)

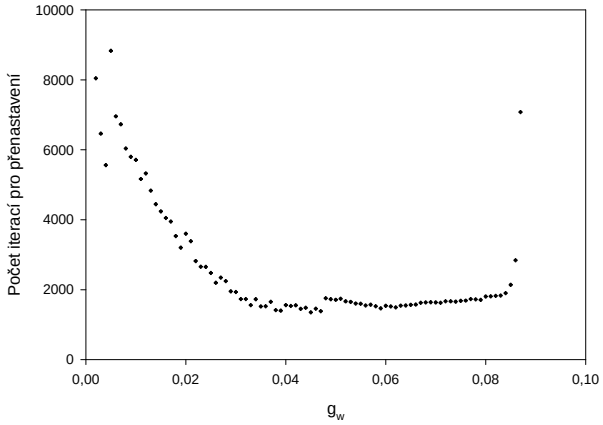



Obr. 3b: Vliv parametru $g_h$ na počet iterací nutný k přeladění z konst hodnoty $0,2$ na konst. hodnotu $0,8$ za podmínek podle (13)

Parametr $g_w$ naopak udává, do jaké míry úspěšná akce ovlivní akce s ní sousedící. Pokud je jeho hodnota nízká, má vliv úspěšná akce na rozložení pravděpodobnosti jen ve svém blízkém okolí. S jeho růstem se rozšiřuje vliv úspěšné akce a tím i rychlost učení (obr. 4a). Významnou nevýhodou vysoké hodnoty parametru $g_w$ je nízká přesnost naučené akce. S rostoucí hodnotou parametru $g_w$ klesá rychlost přeladění (obr. 4b).



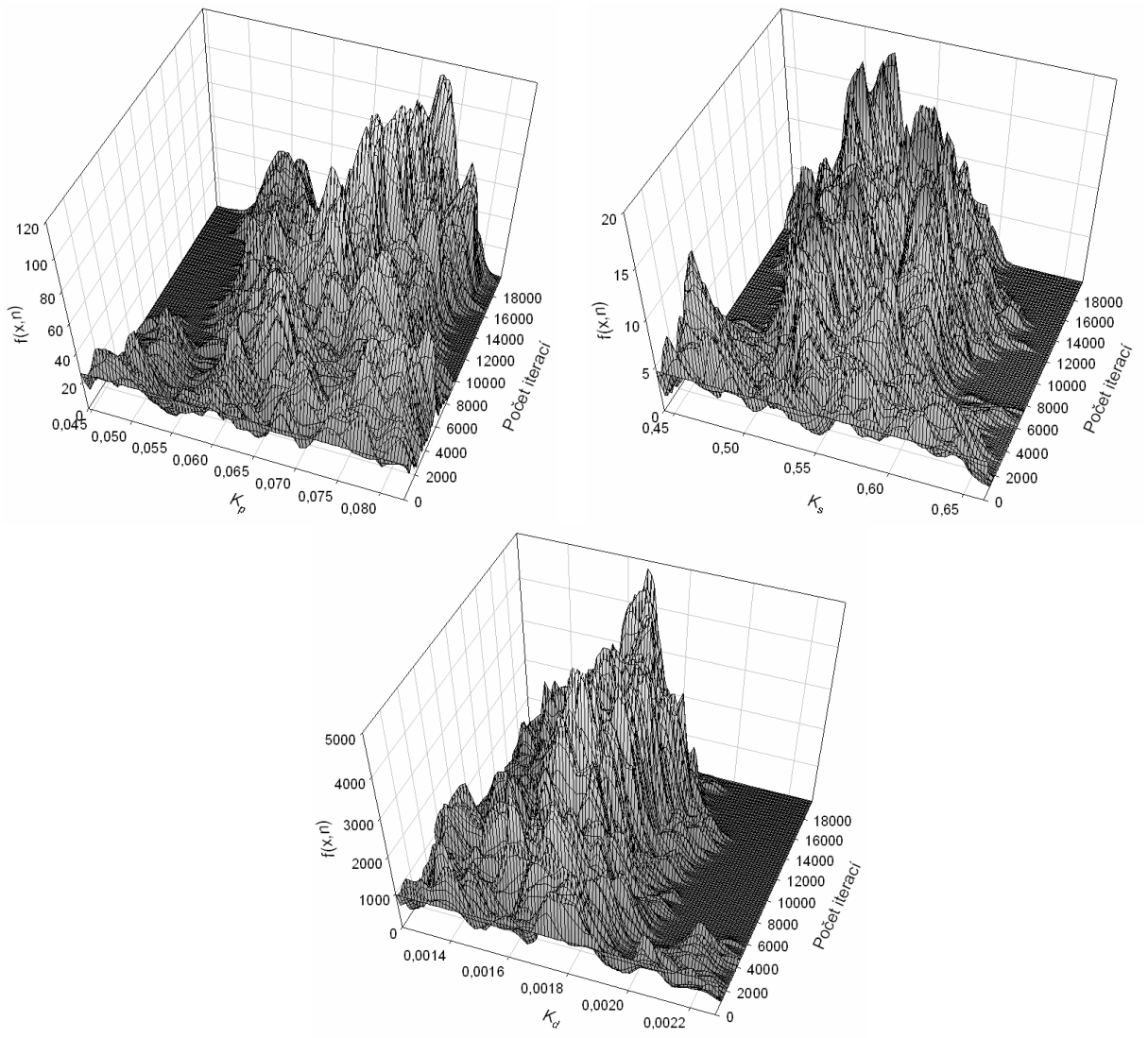

Obr. 4a: Vliv parametru $g_w$ na počet iterací nutný k naladění konst hodnoty $0,2$ za podmínek podle (13)




Obr. 4b: Vliv parametru $g_w$ na počet iterací nutný k přeladění z konst hodnoty $0,2$ na konst. hodnotu $0,8$ za podmínek podle (13)

Z obr. 4a, 4b je patrné, že vhodnou hodnotou je $g_w = 0,03$.

Příklad chování automatů CARLA v extrémních podmínkách je na obr. 5, kde je znázorněn vývoj rozložení $f(x,n)$ pro jednotlivé parametry PSD regulátoru během procesu učení v silně stochastickém prostředí. Simulační experimenty byly prováděny s parametry modelu řízené soustavy podle tab. 2, stochastické prostředí bylo simulováno náhodnými chybami (aditivním šumem) snímače otáček z intervalu $\pm 100\,[\mathrm{min}^{-1}]$. I v takovém prostředí byly automaty CARLA schopny dosáhnout nastavení parametrů PSD regulátoru ($K_d = 1,72.10^{-3}\,[\mathrm{Hz.s}^2]$, $K_s = 5,21.10^{-1}\,[\mathrm{Hz}]$, $K_p = 6,54.10^{-2}\,[\mathrm{Hz.s}]$), se kterými bylo získáno kritérium kvality regulace o 9.8 % lepší než s regulátorem nastaveným Ziegler-Nicholsovou metodikou ($K_d = 1,76.10^{-3}\,[\mathrm{Hz.s}^2]$, $K_s = 5,53.10^{-1}\,[\mathrm{Hz}]$, $K_p = 6,36.10^{-2}\,[\mathrm{Hz.s}]$) za snížení největších překmitů až o 20 % Hodnoty byly naladěny automaty CARLA po $n = 18000$ krocích. V prostředí bez šumu docházelo ke srovnatelnému ustálení hodnot parametrů po již $n = 8000$.




Obr.5: vývoj $K_d$, $K_s$ a $K_p$ během učení, šum snímače otáček $\pm 100\,[\mathrm{min}^{-1}]$

## 6. Závěr

Automaty CARLA jsou schopny zlepšovat funkční charakteristiku řízeného asynchronního elektromotoru i při vysoké úrovni šumu snímače otáček. Tyto vlivy ale přispívaly ke zpomalení učení i při jednoduché struktuře řídicího členu. Zdá se však, že ve srovnání s jinou variantou opakovaně posilovaného učení, tzv. Q-učením (viz. např. Marada, Březina & Singule 2004) je proces učení podstatně rychlejší. Metoda vykazuje vysokou schopnost adaptace parametrů na změnu parametrů prostředí. Tato schopnost však může být na závadu, protože může způsobit i přeladění parametrů PSD regulátoru akumulovanými náhodnými vlivy. Rychlost konvergence i rychlost přelaďování je v širokých mezích ovlivnitelná volbou hodnot dvou parametrů metody ( $g_h$ a. $g_w$ ). V porovnání s Ziegler-Nicholsovým korektorem dosahuje naučený řídicí člen lepší hodnoty kritéria kvality regulace. Vzhledem k pružnosti metody CARLA je pravděpodobné, že bude možno k řízení asynchronního motoru použít i složitější klasická schémata regulátorů, např. regulátor polohy s otáčkovou a proudovou zpětnou vazbou s takto laděnými parametry.

## 7. Poděkování

Tento příspěvek vznikl za podpory výzkumného záměru MSM 262100024 "Výzkum a vývoj mechatronických soustav" a MSM 261100009 "Netradiční metody studia komplexních a neurčitých systémů".

## 8. Reference

Holland, J. (1975) Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor.

Howell, M. N. & Best, M. C. (2000) On-line PID tuning for engine idle-speed control using continuous action reinforcement learning automata. *Control Engineering Practice 8* (2000) 147-154.

Howell, M.N., Frost, G.P., Gordon, T.J. & Wu, Q.H. (1997) Continuous action reinforcement learning applied to vehicle suspension control. *Mechatronics 7* (3), 263–276.

Howell, M.N. & Gordon, T.J. (2001) Continuous action reinforcement learning automata and their application to adaptive digital filter design. *Engineering Applications of Artificial Intelligence 14* (2001), 549–561.

Marada, T., Březina, T. & Singule, V. (2004) Determination of Q-function optimum grid applied on asynchronous electric motor control task, in: *Proc. IM'2004*, Svratka.

Najim, K. & Poznak, A.S. (1994) Learning Automata - Theory and Applications. Pergamon Press, Oxford.

Narendra, K.S. & Thathachar, M.A.L. (1989) Learning Automata: An Introduction, Prentice Hall, London.

Ong, Ch.M. (1998) Dynamic Simulation of Electric Machinery, Prentice Hall, New Jersey.