RIADENIE DYNAMICKÝCH SYSTÉMOV POUŽITÍM Q-UČENIA
CONTROL OF DYNAMICS SYSTEMS BASED ON Q-LEARNING

Anna Filasová, Juraj Klacik, Ján Kašprišin
Katedra kybernetiky a umelej inteligencie, Elektrotechická fakulta, Technická Univerzita Košice, 04200, Košice

Abstrakt Cieľom tohto článku je prezentovať výsledky riešenia problému riadenia nelineárnych systémov použitím algoritmu Q-učenia, ktorý patrí do skupiny algoritmov návrhu adaptívneho kritika (ACD – z.angl. Adaptive Critic Designs). Výhodou tohto prístupu oproti ostatným zo skupiny ACD je, že na nájdenie riadiacej postupnosti regulátora nie je potrebný model systému. Na overenie efektívnosti algoritmu bol použitý model energetického systému.

Summary The purpose of the paper is to present an algorithm for solving nonlinear systems control problem based on Q-learning, which is a model-free approach belonging to the adaptive critic family of designs and its advantage over other algorithms of this family is that it does not need a model of the system. The example application is concerning the model of power system.

1. ÚVOD

V modernej teórii riadenia sú všetky požiadavky na riadenie zhrnuté do kritéria kvality a problém riadenia je transformovaný na optimalizačný problém minimalizácie kritéria kvality. Pri tomto prístupe existujú dva zásadné problémy. Prvým je vhodná voľba kritéria kvality, ktorá by zahrnula všetky naše požiadavky na riadenie a druhým je riešiteľnosť takto formulovaného optimalizačného problému. Najviac používaným kritériom kvality riadenia je kvadratické kritérium, ktoré pre lineárne systémy vedie na lineárny zákon riadenia. V teórii riadenia sa preto na riešenie problému nájdenia optimálneho riadenia lineárnych systémov používa tzv. LQ riadenie (z angl. Linear Quadratic). Pre nelineárne systémy je použitie tohto prístupu veľmi komplikované, a preto je potrebné siahnuť po iných prístupoch, ako je napr. Q-učenie, ktoré na nájdenie optimálneho riadenia v zmysle kvadratického kritéria využíva poznatky umelej inteligencie, konkrétne teóriu umelých neurónových sietí.

2. LINEÁRNE KVADRATICKÉ RIADENIE

Uvažujme lineárny diskrétny časovo optimálny systém ktorého stavový opis je:

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k) \tag{1}$$

$$\mathbf{y}(k) = \mathbf{H}\mathbf{x}(k) + \mathbf{I}\mathbf{u}(k) \tag{2}$$

kde $\mathbf{x}$(k) je vektor stavových veličín, $\mathbf{u}$(k) vektor vstupných a $\mathbf{y}$(k) vektor výstupných veličín v časovom okamihu k.
Úlohou LQ riadenia je nájsť také $\mathbf{u}(k) = -\mathbf{K}(k)\mathbf{x}(k)$ pre systém popísaný rovnicami (1), (2), aby bolo minimalizované kvadratické kritérium (funkcionál)

$$J = \mathbf{x}^T(N)\mathbf{Q}^*\mathbf{x}(N) + \sum_{k=1}^{N-1}(\mathbf{x}^T(k)\mathbf{Q}\mathbf{x}(k) + \mathbf{u}^T(k)\mathbf{R}\mathbf{u}(k)) \tag{3}$$

kde N je prirodzené číslo, $\mathbf{Q}^*$ je symetrická kladne semidefinitná matica, $\mathbf{Q}$ symetrická kladne semidefinitná matica, $\mathbf{R}$ symetrická kladnene definitná matica a $\mathbf{K}(k)$ je postupnosť matíc spätnoväzbových zosilnení.
Úlohou kladnene definitnej matice R vo funkcionáli (3) je zabezpečiť ohraničenie amplitúd prvkov vektora riadiacich veličín $\mathbf{u}$(k) na fyzikálne realizovateľné hodnoty. Zmyslom zavedenia kladne semidefinitnej matice $\mathbf{Q}$ je zabezpečiť konvergenciu amplitúd zložiek stavového vektora do nuly. Optimálne riadenie, okrem minimalizácie funkcionálu (3), musí zabezpečovať asymptotickú stabilitu uzavretého regulačného obvodu, čo možno dosiahnuť vytvorením modifikovaného funkcionálu na základe Ljapunovovej funkcie, pomocou ktorej sa asymptotická stabilita riadenia zabezpečí. Najjednoduchší tvar Ljapunovovej funkcie pre diskrétny lineárny systém je podľa [1], [2]

$$J(\mathbf{x}(k+1)) = \mathbf{x}^T(k+1)\mathbf{P}(k)\mathbf{x}(k+1) \tag{4}$$

Postupnosť matíc zosilnení riadenia $\mathbf{K}(k)$ pre $k = N-1$, $N-2$, ... 0 je možné na základe (3), (4), podľa [1], [2] vypočítať ako

$$\mathbf{K}(k) = (\mathbf{R} + \mathbf{G}^T\mathbf{P}(k)\mathbf{G})^{-1}\mathbf{G}^T\mathbf{P}(k)\mathbf{F} \tag{5}$$

pričom postupnosť matíc $\mathbf{P}(k)$, $k = N-1$, $N-2$, ... 0, pre $\mathbf{P}(N-1) = \mathbf{Q}^*$ je daná riešením Riccatiho rovnice

$$\mathbf{P}(k-1) = \mathbf{Q} + \mathbf{F}^T\mathbf{P}(k)\mathbf{F} - \mathbf{F}^T\mathbf{P}(k)\mathbf{G}\mathbf{K}(k) \tag{6}$$

Vlastnosťou LQ riadenia pre časovo-invariantné systémy a kvadratické funkcinály je, že optimálna postupnosť matíc zosilnení konverguje ku konštantnej matici spätnoväzobných zosilnení $\mathbf{K}$. Riadenie

s ustálenou hodnotou riešenia matice **K** je potom možné zapísať v tvare

$$\mathbf{u}(k) = -\mathbf{K}\mathbf{x}(k) \tag{7}$$

3. METÓDY NÁVRHU ADAPTÍVNEHO KRITIKA

Metódy návrhu adaptívneho kritika (Adaptive critic designs - ACD) predstavujú účinný nástroj na riešenie problému optimalizácie s využitím neurónových sietí. Spájajú v sebe výhody učenia na základe hodnotenia činnosti (Reinforcement learning - RL) a dynamického programovania (DP) za účelom optimalizácie riadenia nelineárnych systémov pracujúcich v prítomnosti porúch a šumov. Ak je proces optimalizácie uvažovaný na časovom intervale < 0; 1 >, potom je možné podľa [2] vážovať optimalizačné kritérium (Bellmanovu rovnicu) v tvare

$$J(k) = \sum_{i=0}^{\infty} \gamma^i U(k+i), \tag{8}$$

kde $\gamma \in$ (0 , 1> je takzvaný znižovací koeficient a U je účelovou funkciou (kvadratickým kritériom).

$$U(k) = \mathbf{x}^{\mathbf{T}}(k)\mathbf{Q}\mathbf{x}(k) + \mathbf{u}^{\mathbf{T}}(k)\mathbf{R}\mathbf{u}(k) \tag{9}$$

pričom $\mathbf{x}(k)$, $\mathbf{u}(k)$ sú vektory stavov, resp. vstupov systému a matice **Q** a **R** sú váhovými maticami stavov, resp. vstupov systému, pričom ich funkcia je rovnaká ako v prípade LQ riadenia.
Ľahko je možné dokázať, že vzťah (8) sa dá nahradiť tzv. Bellmanovou rekurziou takto

$$J(\mathbf{x}(k)) = U(k) + \Delta J(\mathbf{x}(k+1)) \tag{10}$$

kde *J* je funkciou, ktorá v zmysle riadenia odpovedá Ljapunovej funkcii (4). Návrh adaptívneho kritika vo všeobecnosti zahŕňa moduly Aktuátor (Action), Kritik (Critic) a Model, ktoré sú spravidla realizované pomocou umelých neurónových sietí, kde modul Model simuluje cieľ riadenia, Kritik odhaduje hodnoty funkcie *J* z Bellmanovej rovnice (8) dynamického programovania (resp. jej deriváciu, v závislosti od použitej metodiky) vzhľadom na stavy cieľa riadenia. Modul Aktuátor slúži na hľadanie optimálnej postupnosti vektorov riadenia $\mathbf{u}(k)$ pri optimálnej estimácii funkcie J (resp. jej derivácie) z modulu Kritik, pričom riadenie je dané ako

$$\mathbf{u}(k) = \mathbf{A}\mathbf{x}(k), \tag{11}$$

kde A predstavuje váhy siete Aktuátor **A** = -**K**, pričom **K** je matica spätnoväzobných zosilnení.

3.1 Prehľad metód ACD

V roku 1979 existovali tri prístupy k realizácii adaptívneho kritika [3]

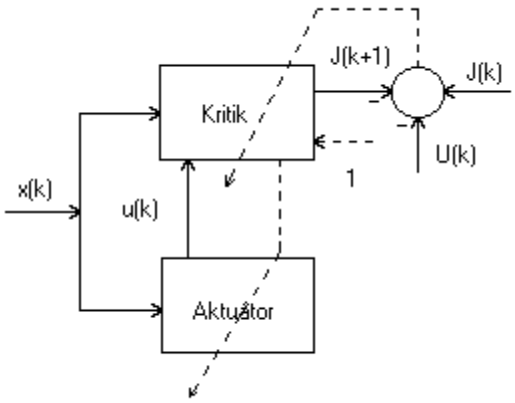
- Heuristické dynamické programovanie (Heuristic dynamic programming – HDP),
- Duálne heuristické programovanie (Dual heuristic programming – DHP),
- Globalizované DHP (Globalized DHP – GDHP).

Tieto prístupy boli navrhnuté tak, že pre riadenie vyžadujú znalosť matematického modelu riadeného systému. V praxi nie je vždy možné pre daný systém nájsť vhodný matematický model. Preto bola navrhnutá metóda Q-učenia, ktorá rieši tento nedostatok.

3.2 Q-učenie

V roku 1989 bola dvoma nezávislými skupinami vedcov navrhnutá modifikácia heuristického dynamického programovania, ktorá bezprostredne spája moduly Aktuátor a Kritik (Model nie je potrebný). Skupina okolo P. Werbosa nazvala túto modifikáciu aktuátorovo-závislým návrhom (action-dependent design – AD) a skupina okolo C. Watkinsa Q-učením (Q-learning). V článku budeme používať pojem Q-učenie.

Na obrázku (Obr. 1) je znázornený princíp adaptácie váh sietí Aktuátor a Kritik v Q-učení.



Obr. 1 Základná schéma Q-učenia
Fig. 1 The Q-learning basic structure

Váhy siete Aktuátor sú adaptované v zmysle hľadania lokálneho extrému funkcie *J* ktorá je výstupom siete Kritik vzhľadom na výstup z modulu Aktuátor $\mathbf{u}(k)$. Výstupnú chybu siete Aktuátor je možné vypočítať ako

$$\mathbf{e}_a(k) = -\frac{\partial J(\mathbf{x}(k))}{\partial \mathbf{u}(k)} \tag{12}$$

Adaptáciu váh siete Aktuátor je potom možné na základe (12) použitím metódy spätného šírenia chyby (z angl. backpropagation – BP) určiť na základe vzťahu

$$\Delta w_{ij}(k) = -\mu \frac{\partial J(\mathbf{x}(k))}{\partial \mathbf{u}(k)} \frac{\partial \mathbf{u}(k)}{\partial w_{ij}(k)} \tag{13}$$

pričom výraz $\partial J(\mathbf{x}(k))/\partial \mathbf{u}(k)$ je získaný priamo spätným šírením signálu cez sieť Kritik smerom k výstupu siete Aktuátor. Keďže sieť Kritik aproximuje funkciu J z Bellmanovej rovnice DP, jej váhy sú adaptované metódou BP, tak, aby bola minimalizovaná výstupná chyba siete, ktorá je na základe (10) daná ako

$$e_k(k) = J(k) - \gamma J(k+1) - U(k) \tag{14}$$

Adaptácia váh siete Kritik je potom vypočítaná ako:

$$\Delta w_{ij}(k) = -\mu (J(k) - \gamma J(k+1) - U(k)) \frac{\partial J(k)}{\partial w_{ij}(k)} \tag{15}$$

3.3 Popis algoritmu Q-učenia

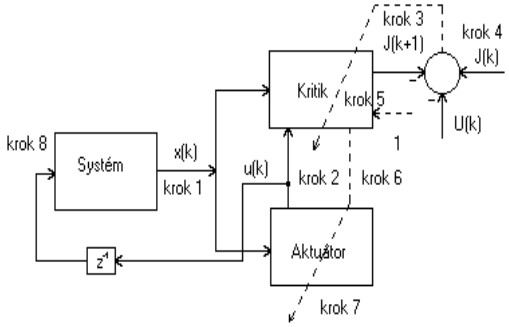
Riešenie problému kvadraticky optimálneho riadenia sústav použitím Q-učenia môžeme zhrnúť podľa [4] do týchto bodov:

1. Zistenie hodnôt jednotlivých koeficientov stavového vektora $\mathbf{x}(k)$ riadeného systému (meraním, v prípade ak to povaha systému dovoľuje, resp. odhadovaním na základe pozorovateľa stavu),
2. Šírením stavového vektora $\mathbf{x}(k)$ cez sieť Aktuátor je na jej výstupe získaný akčný zásah $\mathbf{u}(k)$,
3. Šírením stavového vektora $\mathbf{x}(k)$ a akčného zásahu $\mathbf{u}(k)$ cez sieť Kritik je na jej výstupe získaná hodnota funkcie $J(k+1)$,
4. Na základe (9) je vypočítaná účelová funkcia $U(k)$, ktorá slúži na výpočet chyby siete Kritik $e_k(k)$ podľa (14),
5. Adaptácia váh siete Kritik metódou spätného šírenia chyby,
6. Hodnota parciálnej derivácie $\partial J(\mathbf{x}(k))/\partial \mathbf{u}(k)$ je určená spätným šírením signálu 1 cez neurónovú sieť Kritik smerom k výstupu z Aktuátora, čím sa získa výstupná chyba neurónovej siete Aktuátor $\mathbf{e}_a(k)$,
7. Adaptácia váh siete Aktuátor metódou spätného šírenia chyby,
8. Na vstup do systému je privádzaný akčný zásah $\mathbf{u}(k)$ generovaný sieťou Aktuátor a algoritmus pokračuje bodom 1.

Pre ilustráciu je algoritmus Q-učenia načrtnutý na obrázku (Obr. 2).

4. SIMULÁCIE A VÝSLEDKY

Efektívnosť prezentovaného algoritmu bola overovaná v Matlabe, simuláciami na modeli energetického systému 4. rádu, ktorý je popísaný sústavou nelineárnych diferenciálnych rovníc



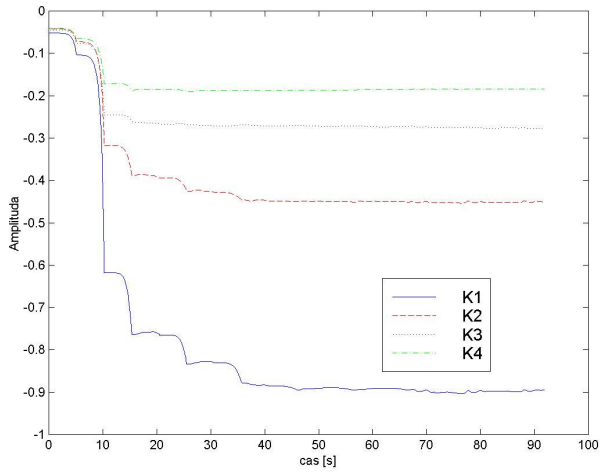
Obr. 2 Schéma algoritmu Q-učenia
Fig. 2 The Q-learning algorithm configuration

$$\begin{aligned}
&\frac{d\delta_m}{dt} = \varpi \\
&M\frac{d\omega}{dt} = -d_m + U + E_m y_m V \sin(\delta - \delta_m - \Theta_m) \\
&K_{qw}\frac{d\delta}{dt} = -K_{qw2}^2 - K_{qw} + E_0^{'} y_0^{'} V \cos(\delta + \Theta_0^{'}) - \\
&-(y_0^{'}\cos\Theta_0 + y_m \cos\Theta_m)V^2 + \\
&+ E_m y_m V \cos(\delta - \delta_m + \Theta_m) \\
&k_4\frac{dV}{dt} = K_{pw}K_{qv}^2 V^2 + (K_{pw}K_{qv} - K_{qw}K_{pv})V + \\
&+\sqrt{K_{qw}^2 + K_{pw}^2}[-E_0^{'} y_0^{'} V \cos(\delta + \Theta_0 - h) - \\
&-E_m y_m V \cos(\delta - \delta_m + \Theta_m - h) + \\
&+(y_0^{'}\cos(\Theta_0 - h) + y_m \cos(\Theta_m - h))V^2]
\end{aligned} \tag{16}$$

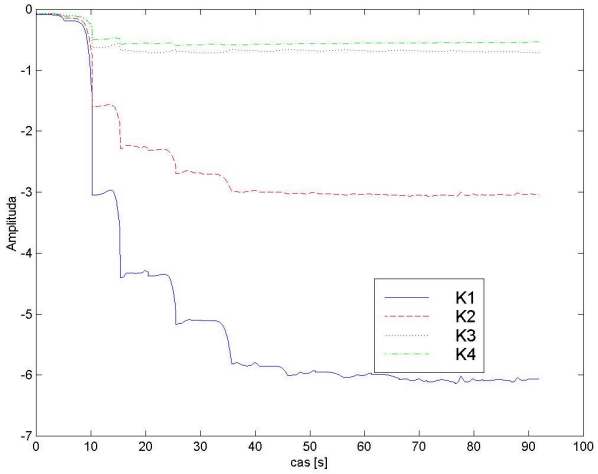
kde $k_4 = TK_{qw}K_{pv}$ a $h = \tan^{-1}(K_{qw}/K_{pw})$.

Nominálne hodnoty koeficientov v rovniciach (16) sú: $K_{pw} = 0.4$, $K_{pv} = 0.3$, $K_{qw} = -0.03$, $K_{qv} = -2.8$, $K_{qv2} = 2.1$, $T = 8.5$, $E_0 = 1.0$, $y'_0 = 8.0$, $\theta'_0 = -12.0$, $E'_0 = 2.5$, $y_m = 5.0$, $\theta_m = -5.0$, $E_m = 1.0$, $M = 0.3$ a $d_m = 0.5$.

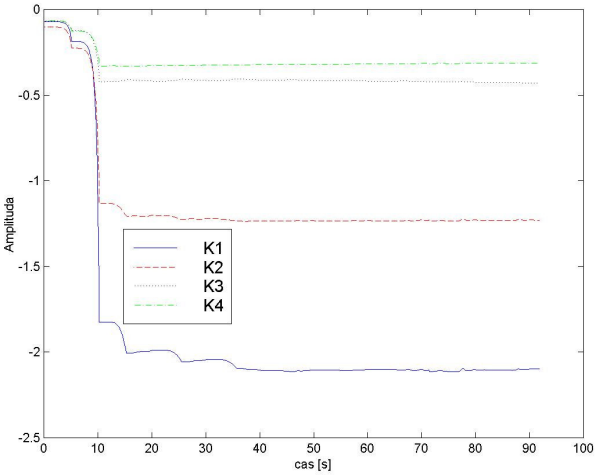
Takto definovaný model systému je v otvorenej slučke nestabilný. Preto bolo potrebné adaptovať váhy sietí Aktuátor a Kritik takým spôsobom, že simulácia bola spúšťaná vždy po piatich sekundách znovu, s novým počiatočným stavovým vektorom. Tým bolo zabezpečené, že stavy neprekračovali v dôsledku nedostatočnej adaptácie váh sietí Aktuátor a Kritik maximálne (fyzikálne realizovateľné) hodnoty. Po ukončení procesu adaptácie váh neurónových sietí, tieto predstavovali optimálny adaptívny regulátor.



Obr. 3 Konvergencia váh siete Aktuátor
Fig.3 Action network weights convergence



Obr. 4 Konvergencia váh siete Aktuátor
Fig. 4 Action network weights convergence



Obr. 5 Konvergencia váh siete Aktuátor
Fig. 5 Action network weights convergence

Množina počiatočných stavových vektorov systému pozostávala zo šiestich prvkov a bola spustená v troch cykloch. Váhy neurónových sietí boli v procese inicializácie nastavené náhodne na hodnoty z intervalu <-0.01; 0.01>. Výsledky simulácií pre rôzne matice **Q** a **R** sú uvedené v tabuľke 1. Proces konvergencie prvkov matice zosilnení, ktorá je aproximovaná sieťou Aktuátor, je pre jednotlivé prípady z tabuľky (Tab. 1) zachytené na obrázkoch (Obr. 3, 4, 5).

*Tab. 1 Výsledky simulácií (**I**- jednotková matica 4x4)*
*Table 1 Results from simulation (**I**-identity matrix 4*4)*

	K	γ_c/γ_a	**Q**	**R**
Obr. 3	0.8954, 0.4508 0.2780, 0.1839	0.0011 0.007	0.003***I**	0.0039
Obr. 4	6.0617, 3.0341 0.7139, 0.5375	0.001 0.001	0.004***I**	0.004
Obr. 5	2.0976, 1.2303 0.4310, 0.3134	0.001 0.003	0.0035***I**	0.0036

Dôležitým prvým krokom pri návrhu regulátora použitím Q-učenia je zvoliť vhodnú účelovú funkciu $U(k)$, a to tak, aby v sebe zahrňovala všetky ciele pre riadený systém. Rovnako ako pri LQ riadení, aj v prípade Q-učenia, je možné vhodnou voľbou matíc **Q** a **R** „ladenie“ regulátora z hľadiska požadovaných vlastností (prechodové a frekvenčné charakteristiky, obmedzenie vstupných veličín a pod.), no na rozdiel od LQ riadenia Q-učenie nezaručuje implicitne stabilitu regulačného obvodu.

Veľkou výhodou tohto prístupu je to, že nie je potrebný model sústavy. Pre návrh regulátora je postačujúce poznať vnútorné stavy riadeného systému (ak nie sú merateľné, je potrebné použiť pozorovateľ stavov), ktoré tvoria spolu s generovaným riadiacim vektorom trénovaciu množinu pre adaptáciu neurónových sietí Aktuátor a Kritik. Pri realizácii Q-učenia sa ako veľmi efektívne ukázalo normovanie n-rozmerného stavového priestoru R^n na interval $< -1, 1>^n$. Všetky stavy, ktoré vstupujú do neurónovej siete Aktuátor sa teda delia ich maximálnou možnou hodnotou. V prípade, že pre neurónovú sieť Aktuátor sa normovanie nerealizuje dochádza k extrémnemu nárastu hodnôt synaptických váh neurónovej sieti, čo následne vedie k zahlteniu a zlyhaniu celého procesu návrhu regulátora.

V prípade ak hodnota stavu prekročí jeho maximálnu možnú hodnotu, čo je v prípade nestabilných systémov pravdepodobné, je potrebné pozastaviť proces učenia, systém nastaviť do inicializačného stavu danej periódy a následne opäť pokračovať v učení, pričom váhy neurónových sietí sa počas inicializácie stavového vektora nemenia. Určenie normovacích koeficientov môže byť analitické na základe maximálnych možných hodnôt stavových premenných, alebo empirické na základe pozorovania správania sa systému. Táto

hodnota do značnej miery ovplyvňuje celý proces návrhu regulátora.

Algoritmus je veľmi citlivý na nastavenie učiacich parametrov sietí Kritik a Aktuátor, ako aj na počiatočnú inicializáciu váh oboch sietí. Aby sa zabezpečila relatívna necitlivosť na počiatočné nastavenia váh oboch sietí, je potrebné zvoliť dostatočne malý interval hodnôt, z ktorého sa budú váhy neurónových sietí náhodne inicializovať. Tento interval však závisí aj od konkrétnej sústavy. V uvažovanom príklade učiace parametre boli stanovené experimentálne.

5. VYHODNOTENIE VÝSLEDKOV

Pre nelineárne sústavy je návrh LQ riadenia veľmi zložitý a vyžaduje linearizáciu sústavy v každom pracovnom bode. Z tohto hľadiska je Q-učenie veľmi vhodnou metódou pre návrh riadenia nelineárnych a ťažko identifikovateľných sústav. Jeho nedostatkom je však relatívne vysoká časová náročnosť, ktorá vyplýva z nutnosti adaptácie váh neurónových sietí iteračným spôsobom. Jeho veľkou výhodou je naopak to, že matematický model riadeného systému nie je potrebný, čo predstavuje značný prínos v prípade riadenia ťažko identifikovateľných sústav. Významným potenciálom Q-učenia, ako aj ostatných metód zo skupiny ACD je ich schopnosť adaptovať sa na zmeny v parametroch regulovanej sústavy.

ZOZNAM POUŽITEJ LITERATÚRY

[1] Krokavec, D., Filasová, A.: Optimálne stochastické systémy. Elfa, Košice 2002, 284s. ISBN 80-89066-52-6.

[2] Filasová, A., Kašprišin, J., Krokavec, D.: Robust LQ control. In Proceedings of the 5th International Scientific – Technical Conference Process Control 2002, 09-12 June, 2002, Kouty nad Desnou, Czech Republic, [CD-ROM] / S. Krejčí, I. Taufer, B. Jakeš, J. Kotyk, J. Macháček, (eds), (Abstract Proceedings, s.46), ISBN 80-7194-452-1.

[3] Kašprišin, J.: Algoritmizácia Kalmanovho estimátora stavu heuristickým dynamickým programovaním. Diplomová práca. KKUI FEI TU, Košice 2000.

[4] Klacik, J.: Riadenie dynamických systémov použitím Q-učenia Diplomová práca. KKUI FEI TU, Košice 2003.

[5] Filasová, A., Kašprišin, J., Krokavec, D.: Stabilization of power transient process. In The 5th International Conference on Control of Power & Heating Systems 2002, 21-22 June, 2002, Zlín, Czech Republic, [CD-ROM] / J. Balátě, B. Chramcov, M. Princ (eds), (Proceedings of Annotations, s.110), ISBN 80-7318-074-X.

[6] Krokavec, D.: Minimal error variance risk-sensitive control. In Proceedings of the 14th International Conference on Process Control ´03, 08-11 June, 2003, Štrbské Pleso, Slovak Republic, [CD-ROM] / J. Mikleš, J. Dvoran, M. Fikar (eds), s.177-1 – 177-6. (Summaries Volume s.93), ISBN 80-227-1902-1.