مجله مهندسی مکانیک مدرس، دی 1395، دوره 16، شماره 10، ص‌ص 241-252

ماهنامه علمی پژوهشی

مهندسی مکانیک مدرس

mme.modares.ac.ir

حل برخط معادله همیلتون-ژاکوبی-بلمن برای سیستم‌های غیرخطی با دینامیک داخلی نامعلوم با استفاده از شبکه‌ی عصبی

آرمان موجودی[1]، مهیار نراقی[2*]، مجتبی مرادی[3]

1- فارغ التحصیل کارشناسی ارشد، مهندسی مکانیک، دانشگاه صنعتی امیرکبیر، تهران
2- دانشیار، مهندسی مکانیک، دانشگاه صنعتی امیرکبیر، تهران
3- دانشجوی دکترا، مهندسی مکانیک، دانشگاه صنعتی امیرکبیر، تهران
* تهران، 15916-34311، naraghi@aut.ac.ir

اطلاعات مقاله

مقاله پژوهشی کامل
دریافت: 01 تیر 1395
پذیرش: 06 شهریور 1395
ارائه در سایت: 24 مهر 1395

کلید واژگان:
معادله همیلتون-ژاکوبی-بلمن
کنترل بهینه
سیستم غیرخطی
شبکه‌ی عصبی
ساختار عملگر- ارزیاب

چکیده

در این مقاله روشی برای حل برخط معادله همیلتون-ژاکوبی-بلمن به منظور طراحی کنترلر بهینه برای سیستم‌های غیرخطی زمان پیوسته ارائه شده است. دیدگاه اساسی در این روش استفاده از تجربیات برای تقویت کنترلر می‌باشد، که با عنوان یادگیری تقویتی معروف است. ابتدا بر اساس ساختار عملگر- ارزیاب و به صورت برخط با استفاده از دو شبکه‌ی عصبی مجزا، معادله همیلتون-ژاکوبی-بلمن به صورت تقریبی حل می‌شود. شبکه‌های عملگر و ارزیاب به ترتیب قانون کنترل بهینه و تابع ارزش بهینه را تخمین می‌زنند. سپس با استفاده از گرادیان نزولی این تخمین‌ها بهبود می‌یابند. از آنجاکه مدل کردن و تعیین مواردی چون اصطکاک و میرایی پیچیده و مشکل می‌باشد، از یک شبکه‌ی عصبی-مقاوم به‌منظور تخمین دینامیک داخلی سیستم استفاده شده است. به این ترتیب ساختار نهایی، عملگر- ارزیاب- شناساگر می‌باشد که با استفاده از آن بدون نیاز به دانستن دینامیک داخلی سیستم، معادله همیلتون-ژاکوبی-بلمن حل و کنترلر بهینه طراحی می‌شود. پایداری روش ارائه شده با استفاده از تابع لیاپانوف اثبات شده است. کارایی روش ارائه شده به صورت عملی برای سیستم خطی موتور DC و با شبیه‌سازی برای یک سیستم غیرخطی نشان داده شده است. نتایج، عملکرد مناسب روش ارائه شده برای حل معادله همیلتون-ژاکوبی-بلمن نشان می‌دهد.

Online solution of the Hamilton–Jacobi–Bellman equation for nonlinear systems with unknown drift dynamics using neural network

Arman Mojoodi, Mahyar Naraghi*, Mojtaba Moradi

Department of Mechanical Engineering, Amirkabir University of Technology, Tehran, Iran
* P.O.B. 15916-34311 Tehran, Iran, naraghi@aut.ac.ir

ARTICLE INFORMATION

Original Research Paper
Received 21 June 2016
Accepted 27 August 2016
Available Online 15 October 2016

Keywords:
HJB equation
Optimal control
Nonlinear system
Neural network
Actor-critic

ABSTRACT

In this paper a method for online solution of the Hamilton-Jacobi–Bellman (HJB) equation is proposed. The method is utilized to design an optimal controller for continuous-time nonlinear systems. The main concept in this approach is using experiences to reinforce the controller, which is called Reinforcement Learning (RL). The online solution is based on the actor-critic (AC) structure where two Neural Networks (NNs) approximately solve the HJB equation. Optimal control and optimal value function are approximated by the actor and the critic, respectively. Then, employing gradient descent algorithm, accuracy of the approximation is improved. Since some items like friction and damping are difficult to model and calculate, a neural-robust identifier is used in conjunction with the AC to approximate drift dynamics. Finally, the Actor-Critic-Identifier (ACI) structure is proposed to solve the HJB equation online with no prior knowledge of drift dynamics. The closed-loop stability of the overall system is assured by the Lyapunov theory employing the direct method. Then the effectiveness of the proposed method is illustrated by experiment for DC motor and simulation for a nonlinear system. Results indicate satisfactory performance of the proposed method to solve the Hamilton-Jacobi-Bellman equation.

1-مقدمه

کنترل بهینه از پایه‌های طراحی سیستم‌های کنترلی مدرن شناخته می‌شود. قانون‌های کنترل بهینه عملکرد مورد نظر برای سیستم را بگونه‌ای تامین می‌کنند که شاخص هزینه کاهش یابد. شاخص هزینه در حقیقت تعادلی میان عملکرد مطلوب و منابع کنترلی موجود می‌باشد [1].

حل مسأله کنترل بهینه نیازمند حل معادله بلمن می‌باشد. در کنترل حلقه بسته، معادله بلمن به معادله همیلتون-ژاکوبی-بلمن که معادله دیفرانسیل جزیی غیرخطی می‌باشد تبدیل می‌شود. اگر سیستم مورد نظر خطی بوده و تابع هزینه به فرم درجه دو باشد، معادله همیلتون-ژاکوبی-بلمن به معادله ریکاتی تبدیل می‌شود. یکی از روش‌های حل معادله ریکاتی توسط کلینمن ارائه شد [2]. در این روش که با نام تکرار قانون[1] شناخته می‌شود،

[1] Policy iteration

برای ارجاع به این مقاله از عبارت ذیل استفاده نمایید:

Please cite this article using:
A. Mojoodi, M. Naraghi, M. Moradi, Online solution of the Hamilton–Jacobi–Bellman equation for nonlinear systems with unknown drift dynamics using neural network, *Modares Mechanical Engineering*, Vol. 16, No. 10, pp. 241-252, 2016 (in Persian)

حل معادله ریکاتی با استفاده از الگوریتمی تکراری بدست می‌آید. با انتخاب ورودی کنترلی اولیه‌ای که سیستم را پایدار کند الگوریتم تکرار قانون با تکرار دو مرحله، همگرایی به پاسخ بهینه را تضمین می‌کند. این دو مرحله عبارتند از:

1- ارزیابی قانون: با حل یک معادله لیاپانوف، هزینه‌ی مربوط به ورودی کنترلی محاسبه می‌شود.

2- بهبود قانون: ورودی جدیدی که نسبت به ورودی قبلی به ورودی بهینه نزدیک‌تر است بدست می‌آید.

اثبات همگرایی روش تکرار قانون برای سیستم‌های غیرخطی توسط ساریدیس و لی بیان شد [3]. هرچند آن‌ها نشان دادند که با استفاده از الگوریتم تکرار قانون می‌توان به کنترل بهینه رسید، اما روشی برای حل معادله لیاپانوف غیرخطی که در هر مرحله از تکرار باید حل می‌شد ارائه نکردند. استفاده از روش تکرار قانون برای حل مسأله کنترل بهینه برای سیستم‌های غیرخطی تا زمانی که بیرد و همکاران [4] با استفاده از تقریب گلرکین روشی برای حل معادله لیاپانوف گفته شده ارائه کردند، در حد یک تئوری بود. اشکال روش بیرد در حل معادله لیاپانوف غیرخطی این بود که در هر مرحله از روش تکرار قانون نیاز به انتگرال‌گیری‌های متعددی است. ابوخلف و لوییس [5] به‌جای استفاده از تقریب گلرکین از شبکه‌ی عصبی در حل معادله لیاپانوف استفاده کردند. استفاده از شبکه‌ی عصبی منجر به سادگی روابط و آسان شدن روش حل شد. بکارگیری دو روش ارائه شده در [4] و [5]، که به ترتیب با استفاده از تقریب گلرکین و شبکه‌ی عصبی معادله همیلتون-ژاکوبی-بلمن را به صورت برون خط حل می‌کنند نیازمند معلوم بودن تمامی دینامیک سیستم است. این دو ویژگی باعث می‌شود که کنترل مبتنی بر این روش‌ها برای سیستم‌های دارای عدم قطعیت به صورت پارامتر متغیر با زمان یا دینامیک مدل نشده واقعا بهینه نباشد.

در کاربردهای مهندسی استفاده از کنترلرهایی که بتوانند با وجود تغییرات سیستم عملکرد مطلوبی داشته باشند ترجیح داده می‌شوند. کنترل تطبیقی امکان طراحی کنترلرهایی را می‌دهد که می‌توانند خود را با تغییرات سیستم وفق دهند. با وجود این مزیت، غالب روش‌های کلاسیک ارائه شده برای طراحی کنترلر تطبیقی از بهینگی فاصله‌ی زیادی دارند. کنترلرهای بهینه-تطبیقی[1] به دسته‌ای از کنترلرها گفته می‌شود که براساس تئوری کنترل بهینه، طراحی می‌شوند اما در طراحی آنها نیازی به معلوم بودن دینامیک کامل سیستم نیست و می‌توانند با وجود عدم قطعیت‌ها باز هم عملکرد بهینه‌ای داشته باشند. استفاده از کنترلرهای بهینه-تطبیقی می‌تواند این امکان را فراهم آورد که از بهینگی و تطبیق‌پذیری به صورت همزمان بهره برد. چگونگی ترکیب این دو مقوله و طراحی کنترلرهای بهینه-تطبیقی با نگاهی به طبیعت و شناخت اصول یادگیری موجودات زنده معلوم شد.

یادگیری تقویتی[2] شاخه‌ای در یادگیری ماشین[3] است که در آن براساس نحوه‌ی یادگیری موجودات زنده، یک عامل، رفتار بهینه را با عمل بر محیط و مشاهده و ارزیابی نتیجه‌ی آن یاد می‌گیرد. در مهندسی کنترل عامل همان کنترلر و محیط، سیستم موردنظر می‌باشد. روش‌های حل یادگیری تقویتی از لحاظ ریاضی به سه دسته تقسیم می‌شوند: برنامه‌ریزی پویا[4]، روش‌های مونت‌کارلو[5] و روش‌های تفاضل زمانی[6] [6]. راه‌حل‌های ارائه شده براساس برنامه‌ریزی پویا نیازمند معلوم بودن مدل سیستم می‌باشند. روش‌های مونت-کارلو گرچه به مدل نیازی ندارند اما استفاده از آن‌ها به صورت برخط بسیار مشکل می‌باشد. از سوی دیگر روش‌های تفاضل زمانی می‌توانند به صورت برخط و بدون مدل استفاده شوند. روش‌های تفاضل زمانی را می‌توان به یادگیری Q[7]، وضعیت-عمل-پاداش-وضعیت-عمل[8] و عملگر-ارزیاب[9] تقسیم کرد. در این میان روش‌های تفاضل زمانی و به‌طور ویژه یادگیری Q و عملگر-ارزیاب تقریبا پر استفاده‌ترین روش‌ها برای حل مسائل یادگیری تقویتی هستند [7].

بسیاری از محققان با استفاده از ساختار عملگر-ارزیاب الگوریتم‌هایی را برای حل برخط مسأله کنترل بهینه ارائه کرده‌اند. در این ساختار، عملگر عملی را انجام می‌دهد و ارزیاب با سنجش نتیجه‌ی آن عمل مقدار نزدیک بودن عمل انجام شده به عمل بهینه را مشخص می‌کند. ارزیابی می‌تواند براساس معیارهایی چون حداقل سوخت، حداقل انرژی، حداقل زمان و موارد دیگر بیان شود. وربی و لوییس [8] با استفاده از دو شبکه‌ی عصبی برای عملگر و ارزیاب، براساس الگوریتم تکرار قانون روشی را برای حل برخط معادله همیلتون-ژاکوبی-بلمن ارائه کردند. در این روش ترکیبی که به صورت زمان پیوسته- گسسته می‌باشد شبکه‌های عملگر و ارزیاب در دو مقیاس زمانی متفاوت عمل می‌کنند؛ به این معنی که شبکه‌ی ارزیاب سریعتر از شبکه‌ی عملگر می‌باشد. ترکیبی بودن این روش به این دلیل است که ورودی کنترلی به صورت پیوسته به سیستم اعمال می‌شود اما در انتهای بازه‌های زمانی گسسته، وزن‌های کنترلر بهنگام‌سازی می‌شوند. الگوریتم تکرار قانون همزمان که در آن ارزیابی و بهبود ورودی کنترلی با هم انجام می‌شود را ومووداکیس و لوییس پیشنهاد دادند [9]. مزیت این روش نسبت به روش وربی و لوویس [8] در یادگیری همزمان دو شبکه‌ی عصبی بود.

نحوه‌ی تعیین قانون بهینه در روش عملگر-ارزیاب که براساس آن، عامل در هر وضعیت تعیین می‌کند که چه رفتاری باید انجام شود اساس تقسیم-بندی دیگری برای روش‌های حل به صورت زیر می‌باشد [10]:

- الگوریتم‌های تکرار ارزش[10]: در این الگوریتم‌ها تابع ارزش بهینه بدست می‌آید و براساس آن قانون بهینه تعیین می‌شود. این الگوریتم‌ها می-توانند وابسته به مدل باشند و یا نباشند.
- الگوریتم‌های تکرار قانون[11]: در این الگوریتم‌ها تابع ارزش با ارزیابی قانون فعلی بدست می‌آید و با استفاده از تابع ارزش جدید، قانون دیگری بدست می‌آید که نسبت به قانون قبلی به بهینگی نزدیک‌تر است. این الگوریتم‌ها نیز می‌توانند وابسته به مدل باشند و یا نباشند. به صورت کلی در این الگوریتم‌ها قانون اولیه باید پایدار کننده باشد.
- الگوریتم‌های جستجوی قانون[12]: در این الگوریتم‌ها قانون بهینه به صورت مستقیم و با استفاده از تکنیک‌های بهینه‌سازی تعیین می‌شود. این الگوریتم‌ها می‌توانند براساس گرادیان باشند و یا نباشند.

روش‌های فوق به معلوم بودن دینامیک سیستم به صورت کامل نیاز دارند. در این مقاله با تعمیم روش ارائه شده توسط ومووداکیس و لوویس [9]، برای یک سیستم غیرخطی، معادله همیلتون-ژاکوبی-بلمن بدون نیاز به معلوم بودن دینامیک داخلی سیستم و به صورت برخط حل می‌شود. برای تخمین دینامیک سیستم از یک شبکه‌ی عصبی-مقاوم استفاده شده است. این شبکه

[1] Optimal adaptive controller
[2] Reinforcement learning
[3] Machine Learning
[4] Dynamic Programming
[5] Monte Carlo
[6] Temporal Difference
[7] Q-Learning
[8] State-Action-Reward-State-Action
[9] Actor-Critic
[10] Value Iteration (VI)
[11] Policy Iteration (PI)
[12] Policy Search (PS)

برگرفته از کارهای ژیان و همکاران [11]، پاتره و همکاران [12] و داپری و همکاران [13] می‌باشد. توجه شود که در [13] ابتدا فرض شده است که معادله دینامیکی سیستم بصورت اویلر-لاگرانژ است و سپس با حل معادله همیلتون-ژاکوبی-بلمن برای سیستم ساده شده کنترلر بهینه طراحی شده است. به همین دلیل روش ارائه شده عمومیت ندارد.

همچنین این مقاله نسبت به [9] این برتری را دارد که روش پیشنهادی بصورت عملی اجرا شده است و به این ترتیب می‌توان کارایی روش پیشنهادی را برای کاربردهای عملی مورد ارزیابی قرار داد.

2- مسأله کنترل بهینه و الگوریتم تکرار قانون

2-1-کنترل بهینه و معادله زمان پیوسته‌ی همیلتون-ژاکوبی-بلمن

سیستم غیرخطی افاین[1] در ورودی مطابق رابطه (1) در نظر گرفته می‌شود.

$$\dot{x} = f(x) + g(x)u \tag{1}$$

که در آن $x \in \Re^n$، $f(x) \in \Re^n$ و $g(x) \in \Re^{n\times m}$. هم چنین ورودی کنترلی، $u(t) \in \Re^m$ می‌باشد. منظور از افاین بودن در ورودی آن است که ورودی به صورت خطی در معادلات دینامیکی ظاهر می‌شود. فرضیات درنظر گرفته شده برای سیستم فوق عبارتند از:

فرض اول: $f(0) = 0$ که به‌معنای قرار گرفتن نقطه‌ی تعادل در مبدأ می‌باشد.

فرض دوم: $f(x) + g(x)u$ بر روی مجموعه‌ی شامل مبدا $\Omega \subseteq \Re^n$، لیپشیتز[2] پیوسته می‌باشد. یعنی مشتق تابع در تمامی دامنه آن کران‌دار است.

فرض سوم: سیستم بر روی مجموعه‌ی Ω پایدارپذیر می‌باشد. منظور از پایدارپذیری وجود تابع کنترل پیوسته‌ی $u(t) \in U$ می‌باشد به نحوی که سیستم مدنظر روی Ω پایدار مجانبی باشد.

هدف یافتن ورودی مناسبی است که تابع هزینه‌ی بیان شده در رابطه (2) را کمینه کند.

$$V(x_0) = \int_0^\infty (Q(x) + R(u))dt \tag{2}$$

در رابطه (2)، $Q(x)$ و $R(u)$ توابع مثبت معین بر روی Ω می‌باشند. همچنین $\forall x \neq 0\ Q(x) > 0$ و $x = 0 \Rightarrow Q(x) = 0$. یک انتخاب معمول برای $R(u)$ به صورت $R(u) = u^{\mathrm{T}}Ru$ می‌باشد که $R \in \Re^{m\times m}$ و در روابط این مقاله $R(u)$ به همین صورت فرض شده است. توجه شود که ورودی u علاوه بر پایدارسازی سیستم روی Ω، باید تابع هزینه‌ی محدودی را بدست دهد. به این ورودی‌های کنترلی، مجاز[3] گویند.

تعریف 1: به ورودی کنترلی u، نسبت به رابطه (1) و بر روی Ω، کنترل مجاز گفته می‌شود و با $u \in \Psi(\Omega)$ نشان داده می‌شود اگر شرایط زیر را داشته باشد؛

u بر روی Ω پیوسته باشد، رابطه $u(0) = 0$ برقرار باشد، رابطه (2) را بر روی Ω پایدار کند و همچنین برای $\forall x_0 \in \Omega$، $V(x_0)$ محدود باشد [5]. Ψ فضای ورودی‌های مجاز را بیان می‌کند.

برای محاسبه‌ی معادله (2) نیاز است که پاسخ سیستم، به صورت کامل مشخص باشد که در حالت کلی چنین پاسخی وجود ندارد. برای مستقل ساختن این معادله از پاسخ سیستم، از این معادله باید مشتق گرفت که در ادامه، این روند توضیح داده می‌شود.

معادله (2) را می‌توان مطابق رابطه (3) بسط داد.

$$\begin{aligned} V(x_0) &= \int_0^T (Q(x) + R(u))dt + \int_T^\infty (Q(x) + R(u))dt \\ &= \int_0^T (Q(x) + R(u))dt + V(x(T)) \end{aligned} \tag{3}$$

توجه شود که تابع ارزش بیان شده در رابطه (3) را می‌توان به صورت ترکیبی از هزینه تا زمان T و مقدار در آن موقعیت نوشت. در صورتی که تابع ارزش V در نقطه‌ی x_0 مشتق‌پذیر باشد، می‌توان معادله (3) را به صورت رابطه (4) بازنویسی کرد.

$$\begin{aligned} &\lim_{T\to 0} \frac{V(x_0) - V(x(T))}{T} = \lim_{T\to 0} \frac{1}{T}\int_0^T (Q(x) + R(u))dt \\ &\dot{V} = \frac{dV}{dt} = \frac{dV}{dx}\frac{dx}{dt} = V_x(f + gu) = -Q(x) - R(u) \end{aligned} \tag{4}$$

توجه شود که در این مقاله مشتق اسکالر نسبت به بردار به صورت بردار سطری بیان شده است بنابراین مشتق تابع هزینه برداری سطری است.

$$V_x = \begin{bmatrix} \frac{\partial V}{\partial x_1} & \frac{\partial V}{\partial x_2} & \cdots & \frac{\partial V}{\partial x_n} \end{bmatrix}$$

معادله (4) نمایش دیفرانسیلی رابطه (2) است و معادله لیاپانوف غیرخطی را می‌توان به صورت رابطه (5) تعریف کرد.

$$\mathrm{LE}(V, u) \triangleq V_x(f + gu) + Q(x) + R(u) = 0, V(0) = 0 \tag{5}$$

به معادله (5) معادله همیلتون-ژاکوبی-بلمن تعمیم‌یافته گویند.

تعریف 2: برای کنترل مجاز $u \in \Psi(\Omega)$، اگر معادله (5) برقرار باشد تابع $V: \Omega \to \Re$ معادله تعمیم‌یافته‌ی همیلتون-ژاکوبی-بلمن را که به صورت $\mathrm{GHJB}(V; u) = 0$ نوشته می‌شود ارضا می‌کند [4]. در افق بی‌نهایت مقدار V ثابت و پیوسته خواهد بود و می‌توان همیلتونین را مطابق رابطه (6) کمینه کرد.

$$\begin{aligned} &H(x, u, V_x) = V_x(f + gu) + Q(x) + R(u) \\ &\hat{u}(x) = \arg \min_{u\in\Psi(\Omega)} \{V_x(f + gu) + Q(x) + R(u)\} \\ &\frac{\partial H}{\partial u} = 0 \Rightarrow \hat{u}(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}V_x^{\mathrm{T}} \end{aligned} \tag{6}$$

هزینه‌ی $\hat{u}$ با حل معادله $\mathrm{GHJB}(\hat{V}; \hat{u}) = 0$ بدست می‌آید و در [3] نشان داده شده است که برای هر حالت پایدارپذیر، $\hat{V}(x) \leq V(x)$ و با تکرار این فرآیند تابع ارزش به تابع ارزش بهینه (حل معادله همیلتون-ژاکوبی-بلمن) همگرا می‌شود.

با جایگذاری کنترل بهینه در معادله لیاپانوف، معادله همیلتون-ژاکوبی-بلمن بدست می‌آید. قانون کنترلی بهینه در رابطه (7) تعریف شده است.

$$u^* = -\frac{1}{2}R^{-1}g^{\mathrm{T}}V_x^{*\mathrm{T}} \tag{7}$$

در رابطه (7)، V_x^*، تابع ارزش بهینه می‌باشد که از حل معادله همیلتون-ژاکوبی-بلمن بدست می‌آید که در رابطه (8) تعریف شده است.

$$\mathrm{HJB}(V^*) \triangleq V_x^* f + Q(x) - \frac{1}{4}V_x^* g R^{-1} g^{\mathrm{T}} V_x^{*\mathrm{T}} = 0, V^*(0) = 0 \tag{8}$$

2-2-روش تکرار قانون

در معادله لیاپانوف، مشتق تابع هزینه به صورت خطی و در معادله همیلتون-ژاکوبی-بلمن مشتق تابع ارزش به صورت غیرخطی ظاهر می‌شود. حل معادله لیاپانوف برای تابع هزینه نیازمند حل معادله دیفرانسیل جزیی خطی است، درحالیکه حل معادله همیلتون-ژاکوبی-بلمن از حل معادله دیفرانسیل جزیی غیرخطی بدست می‌آید که ممکن است امکان‌پذیر نباشد. این امر دلیل استفاده از تکنیک تکرار قانون برای حل معادله همیلتون-ژاکوبی-بلمن است

[1] Affine
[2] Lipchitz
[3] Admissible

که در [3] اثبات شده است. در روش تکرار قانون بجای حل مستقیم معادله همیلتون-ژاکوبی-بلمن و یافتن V^*، با شروع از یک کنترل مجاز و با تکرار دو مرحله‌ی زیر، قانون کنترلی بهینه بدست می‌آید:

1-برای ورودی کنترلی $u^i(x)$، $V^i(x)$ برابر رابطه (9) است.

$$V_x^{(i)}(f + gu^{(i)}) + Q(x) + R(u^{(i)}) = 0, V^{(i)}(0) = 0 \quad (9)$$

2-با استفاده از $V^i(x)$ بدست آمده، ورودی کنترلی مطابق رابطه (10) بهبود می‌یابد.

$$u^{(i+1)}(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}V_x^{\mathrm{T}(i)} \quad (10)$$

3-2-تخمین تابع ارزش

روش تکرار قانون استاندارد که توضیح داده شد از توالی دو مرحله‌ی بیان شده در روابط (9) و (10) تشکیل شده است که به ترتیب تابع ارزش و قانون کنترلی را بهنگام‌سازی می‌کنند. در ادامه از دو رابطه گفته شده برای طراحی دو شبکه‌ی عصبی استفاده می‌شود که هر کدام از آن‌ها نقش یکی از این معادلات را ایفا می‌کنند. بهنگام‌سازی وزن‌های این شبکه‌ها معادل است با به‌هنگام‌سازی ارزش و قانون.

قابلیت شبکه‌های عصبی در تخمین توابع هموار بر روی مجموعه‌های فشرده امری شناخته شده است [14]. منظور از مجموعه فشرده در فضای n-بعدی به معنای بسته و کراندار بودن آن مجموعه است. به منظور استفاده از شبکه‌ی عصبی برای حل مسأله کنترل بهینه فرض می‌شود که تابع هزینه‌ی بهینه پیوسته بوده و بر روی مجموعه‌ای فشرده تعریف شده است. با استفاده از فرضیاتی می‌توان از شبکه‌ی عصبی برای تقریب آن استفاده کرد [9]:

فرض اول: تابع پیوسته‌ی $\Gamma: \mathbb{S} \to \Re^n$ را درنظر بگیرید که در آن $\mathbb{S}$ مجموعه‌ی فشرده و متصل ساده (یعنی دارای نواحی تعریف نشده در بازه نباشد) است؛ آنگاه وزن‌های ایده‌آل W و K وجود دارند به گونه‌ای که تابع را بتوان با شبکه‌ی عصبی بیان شده در رابطه (11) نشان داد.

$$\Gamma(x) = W^{\mathrm{T}}\sigma(K^{\mathrm{T}}x) + \varepsilon(x) \quad (11)$$

در رابطه (11)، $\sigma(\cdot)$ تابع فعالیت غیرخطی و $\varepsilon(x)$ خطای بازسازی تابع می‌باشد.

فرض دوم: وزن‌های ایده‌آل شبکه‌ی عصبی، محدود به ثابت‌های مثبت معلوم می‌باشند؛ یعنی $\|W\| \le \overline{W}$ و $\|K\| \le \overline{K}$.

فرض سوم: تابع فعالیت شبکه‌ی عصبی $\sigma(\cdot)$ و مشتق آن نسبت به ورودی‌های خود، $\sigma'(\cdot)$، محدود می‌باشد.

فرض چهارم: خطای بازسازی تابع و مشتق آن نسبت به ورودی‌های آن محدود می‌باشد ($\|\varepsilon(\cdot)\| \le \overline{\varepsilon}$ و $\|\varepsilon'(\cdot)\| \le \overline{\varepsilon'}$).

با استفاده از این فرضیات و جایگذاری تابع هزینه به صورت شبکه‌ی عصبی و قانون کنترلی بهینه بر آن اساس، رابطه (12) بدست می‌آید.

$$V^*(x) = W^{\mathrm{T}}\phi(x) + \varepsilon(x)$$
$$u^*(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)(\nabla\phi(x)^{\mathrm{T}}W + \varepsilon'(x)^{\mathrm{T}}) \quad (12)$$

در رابطه (12)، $W \in \Re^N$ وزن‌های ایده‌آل شبکه می‌باشند که مقدار آن‌ها معلوم نیست و N بیانگر تعداد نورون‌ها می‌باشد. توابع فعالیت $\phi(x) \triangleq [\phi_1(x)\ \phi_2(x)\ \cdots\ \phi_N(x)]^{\mathrm{T}} \in \Re^N$ می‌باشند و مشتق آن‌ها با $\nabla\phi(x) \triangleq \partial\phi/\partial x \in \Re^{N\times n}$ نشان داده شده است. توابع فعالیت به‌گونه‌ای هستند که $\forall i \in [1\ N]$ دو عبارت $\phi_i(0) = 0$ و $\nabla\phi(0) = 0$ برقرار باشند. همچنین در رابطه (12)، $\varepsilon(\cdot) \in \Re$ خطای بازسازی تابع می‌باشد.

توجه شود که شرط فوق را می‌توان برای هر تابع دلخواه در نظر گرفت اگر $\phi(x)$ مطابق رابطه (13) تعریف شود.

$$\phi_i(x) = [f(x) - f(0)]^2 \quad (13)$$

3-به‌هنگام‌سازی وزن‌های شبکه‌های عصبی

باتوجه به اینکه وزن‌های ایده‌آل شبکه‌های عصبی در دسترس نمی‌باشد بنابراین از خطای بازسازی تابع ارزش صرف‌نظر شده و با تقریب وزن‌های تابع ارزش، رابطه (12) به صورت رابطه (14) بازنویسی می‌شود.

$$\hat{V}(x) = \hat{W}_c^{\mathrm{T}}\phi(x);\ \ \hat{u}(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\phi'^{\mathrm{T}}(x)\hat{W}_a \quad (14)$$

در رابطه (14) وزن‌های شبکه‌ها با اندیس c و a مشخص شده‌اند که به ترتیب بیانگر شبکه‌های ارزیاب و عملگر می‌باشند. مقدار این وزن‌ها با استفاده از قوانین به‌هنگام‌سازی که در ادامه استخراج می‌شوند به صورت برخط بدست می‌آیند.

3-1- شبکه‌ی ارزیاب

هدف شبکه‌ی ارزیاب تقریب زدن تابع ارزش بهینه می‌باشد. در صورتیکه وزن‌های ایده‌آل معلوم باشند می‌توان از رابطه (12) تابع هزینه بهینه را تقریب زد. با استفاده از این تقریب برای یک ورودی کنترلی ثابت u تابع لیاپانوف غیرخطی (6) با استفاده از شبکه عصبی به صورت رابطه (15) بیان می‌شود.

$$\mathrm{H}(x, W, u) = x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru + W^{\mathrm{T}}\nabla\phi(f + gu) = \varepsilon_{\mathrm{H}} \quad (15)$$

خطای باقیمانده در رابطه (15)، از خطای بازسازی تابع، رابطه (16)، حاصل شده است.

$$\varepsilon_{\mathrm{H}} = -(\nabla\varepsilon)^{\mathrm{T}}(f + gu) \quad (16)$$

از آنجاکه مقدار وزن‌های ایده‌آل معلوم نیست از رابطه (14) مقادیر تقریبی جایگزین می‌گردند. به‌منظور استفاده از این رابطه به دو طرف رابطه (15) عبارت $\hat{W}_c^{\mathrm{T}}\nabla\phi(f + gu)$ اضافه می‌شود که حاصل آن در رابطه (17) بیان شده است.

$$\mathrm{H}(x, \hat{W}_c, u) = x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru + \hat{W}_c^{\mathrm{T}}\nabla\phi(f + gu) = e \quad (17)$$

خطای بیان شده در رابطه (17)، e، در رابطه (18) تعریف شده است.

$$e = -\tilde{W}_c^{\mathrm{T}}\nabla\phi(f + gu) + \varepsilon_{\mathrm{H}} \quad (18)$$

که $\tilde{W}_c = W_c - \hat{W}_c$ بیانگر خطای میان وزن‌های ایده‌آل و تقریبی می‌باشد.

مطلوب است $\hat{W}_c$ به‌گونه‌ای تعیین شود که مربع خطای بیان شده در رابطه (17)، $E = \frac{1}{2}e^{\mathrm{T}}e$، کمینه شود. درصورت کمینه شدن مربع خطا دو مسأله رخ می‌دهد؛ $\hat{W}_c \to W$ و $e \to \varepsilon_{\mathrm{H}}$. برای کم کردن مربع خطا از الگوریتم کاهش گرادیان استفاده می‌شود. جهت عکس گرادیان راستای بیشترین کاهش را نشان می‌دهد و حرکت در این جهت باعث می‌شود به سمت کمینه حرکت کرد. با توجه به این بحث، قانون به‌هنگام‌سازی به صورت رابطه (19) قابل بیان است.

$$\dot{\hat{W}}_c = -a\frac{\partial E}{\partial \hat{W}_c} = -a\ e\frac{\partial e}{\partial \hat{W}_c} = -a\,e\,\nabla\phi(f + gu) = -a\,e\,\sigma_c \quad (19)$$

که در آن a ضریب یادگیری می‌باشد. همچنین $\sigma_c = \nabla\phi(f + gu)$. با جایگذاری رابطه (17) در رابطه (19) و با نرمال کردن σ_c شکل نهایی قانون به‌هنگام‌سازی برابر با رابطه (20) است.

$$\dot{\hat{W}}_c = -a\ \frac{\sigma_c}{(\sigma_c^{\mathrm{T}}\sigma_c + 1)^2}[\sigma_c^{\mathrm{T}}\hat{W}_c + Q(x) + u^{\mathrm{T}}Ru] \quad (20)$$

برای محاسبه‌ی دینامیک خطای تخمین وزن‌های ارزیاب، با استفاده از رابطه (15)، رابطه (21) بدست می آید.

$$x^{\mathrm{T}}Qx + u^{\mathrm{T}}Ru = W_c^{\mathrm{T}}\nabla\phi(f + gu) + \varepsilon_{\mathrm{H}} \quad (21)$$

با جایگذاری رابطه (21) در رابطه (20) دینامیک خطای تخمین مطابق با رابطه (22) بدست می‌آید.

$$\dot{\widetilde{W}}_c = -a\,\bar{\sigma}_c\bar{\sigma}_c^T\widetilde{W}_c + a\bar{\sigma}_c\frac{\varepsilon_{\mathrm{H}}}{m_s} \tag{22}$$

که در آن $m_s = \sigma_c^{\mathrm{T}}\sigma_c$ و $\bar{\sigma}_c = \sigma_c/m_s$. هرچند با استفاده از رابطه (20) وزن-های شبکه‌ی ارزیاب به گونه‌ای تغییر می‌کنند که مربع خطا کاهش یابد اما مشخص نیست که همگرایی وزن‌ها چه زمانی رخ می‌دهد. به‌منظور اثبات همگرایی وزن‌ها به وزن‌های ایده‌آل، $\widehat{W}_c \to W_c$، فرض تحریک دائمی و لم-های زیر مورد نیاز می‌باشند.

فرض تحریک دائمی: فرض شود که سیگنال $\bar{\sigma}_c$ بر روی بازه‌ی $[t, t+T]$ دائما تحریک می‌شود. این مسأله به این معنی است که ثابت‌های $\beta_1 > 0$، $\beta_2 > 0$ و $T > 0$ وجود دارند بطوریکه برای تمامی t ها رابطه (23) برقرار است.

$$\beta_1 I \le S_0 \equiv \int_t^{t+T} \bar{\sigma}_c(\tau)\bar{\sigma}_c^{\mathrm{T}}(\tau)d\tau \le \beta_2 I \tag{23}$$

در کنترل تطبیقی به‌منظور شناسایی سیستم فرض تحریک دائمی نیاز است. در اینجا نیز به دلیل آنکه هدف شناسایی پارامترهای ارزیاب می‌باشد این فرض نیاز است.

لم 1: دینامیک خطای وزن‌های شبکه‌ی ارزیاب را با خروجی بیان شده در رابطه (24) در نظر بگیرید.

$$\begin{aligned}\dot{\widetilde{W}}_c &= -a\,\bar{\sigma}_c\bar{\sigma}_c^{\mathrm{T}}\widetilde{W}_c + a\bar{\sigma}_c\frac{\varepsilon_H}{m_s}\\ y &= \bar{\sigma}_c^{\mathrm{T}}\widetilde{W}_c\end{aligned} \tag{24}$$

شرط تحریک دائمی بیان شده در رابطه (23) برای این سیستم معادل است با مشاهده‌پذیری کامل یکنواخت[1] سیستم. به این معنا که ثابت‌های $\beta_3 > 0$، $\beta_4 > 0$ و $T > 0$ وجود دارند بطوری که برای تمامی t ها رابطه (25) برقرار باشد.

$$\beta_3 I \le S_1 \equiv \int_t^{t+T} \Phi^{\mathrm{T}}(\tau,t)\bar{\sigma}_c(\tau)\bar{\sigma}_c^{\mathrm{T}}(\tau)\Phi(\tau,t)d\tau \le \beta_4 I \tag{25}$$

که در آن $t_0 \le t_1$ و $\Phi(t_1,t_0)$ ماتریس انتقال حالت سیستم (24) می‌باشد.

اثبات. در صورتیکه $u = -y + \varepsilon_H/m_s$ ورودی سیستم باشد سیستم (24) و سیستم $\dot{\widetilde{W}}_c = a\,\bar{\sigma}_c u$ با خروجی $y = \bar{\sigma}_c^{\mathrm{T}}\widetilde{W}_c$ برابر هستند. توجه شود که رابطه (23) گرامیان مشاهده‌پذیری[2] برای این سیستم می‌باشد.

اهمیت مشاهده‌پذیری کامل یکنواخت این است که از ورودی و خروجی محدود، محدود بودن متغیر وضعیت نتیجه می‌شود [15].

لم 2: دینامیک خطای (24) را در نظر بگیرید. در صورتیکه سیگنال $\bar{\sigma}_c$ دائما تحریک شود، می‌توان نتیجه گرفت که:

- سیستم (24) پایدار مجانبی است و اگر $\varepsilon = 0$ آنگاه $\|\widetilde{W}(kT)\| \le \exp(-\alpha kT)\|\widetilde{W}(0)\|$ خواهد بود و α مطابق رابطه (26) است.

$$\alpha = -\frac{1}{T}\ln(\sqrt{1-2a\beta_3}) \tag{26}$$

- در صورتی‌که $\|\varepsilon_{\mathrm{H}}\| \le \varepsilon_{\max}$ و $\|y\| \le y_{\max}$ آنگاه $\|\widetilde{W}_c\|$ به صورت نمایی به مجموعه‌ی باقیمانده‌ی[3] بیان شده در رابطه (27) همگرا می‌شود.

$$\widetilde{W}_c(t) \le \frac{\sqrt{\beta_2 T}}{\beta_1}\{[y_{\max} + \delta\beta_2 a(\varepsilon_{\max} + y_{\max})]\} \tag{27}$$

در اینجا δ ثابت مثبت از مرتبه‌ی 1 می‌باشد.

قضیه نشان می‌دهد که قانون به‌هنگام‌سازی (20) در صورت برقراری شرط تحریک دائمی، وزن‌های $\widehat{W}_c$ را به وزن‌های نامعلوم W_c همگرا می‌کند که با استفاده از آن‌ها برای کنترل $u(t)$ داده شده، معادله (15) حل می‌شود.

قضیه 1: فرض کنید که $u(t)$ هر قانون کنترلی محدود مجاز باشد. در صورتیکه به‌هنگام‌سازی شبکه‌ی عصبی ارزیاب با استفاده از (20) انجام شود و $\bar{\sigma}_c$ دائما تحریک شود، با فرض اینکه خطای باقیمانده‌ی رابطه (15) محدود باشد، $\|\varepsilon_{\mathrm{H}}\| \le \varepsilon_{\max}$، خطای پارامترهای شبکه‌ی ارزیاب با فاکتور کاهشی نشان داده شده در رابطه (26) به مجموعه‌ی مانده‌ی تعریف شده در (28) همگرا می‌شود.

$$\widetilde{W}_c(t) \le \frac{\sqrt{\beta_2 T}}{\beta_1}\{[1 + 2\delta\beta_2 a]\varepsilon_{\max}\} \tag{28}$$

اثبات در [9] ارائه شده است.

2-3- شبکه‌ی عملگر

باتوجه به رابطه (12) مشخص است که شبکه‌ی ارزیاب به تنهایی برای حل مسأله بهینه کافی است؛ چرا که قانون کنترلی بهینه از گرادیان تابع هزینه‌ی بهینه بدست می‌آید و با گرادیان گرفتن، وزن‌های شبکه تغییری نمی‌کنند. درنتیجه قانون کنترلی به صورت رابطه (29) است.

$$u(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla\phi^{\mathrm{T}}(x)W_c \tag{29}$$

از آنجا که مقادیر وزن‌های ایده‌آل شبکه‌ی ارزیاب معلوم نمی‌باشند، همچنین به‌منظور امکان ارائه‌ی اثبات پایداری، شبکه‌ی عصبی عملگر برای تخمین قانون کنترلی به صورت رابطه (30) ارائه می‌شود.

$$\hat{u}(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla\phi^{\mathrm{T}}(x)\widehat{W}_a \tag{30}$$

در رابطه (30)، $\widehat{W}_a$ بیانگر تخمینی از وزن‌های ایده‌آل W_c می‌باشد. همچنین خطای تخمین شبکه‌ی عصبی مطابق با رابطه (31) است.

$$\widetilde{W}_a = W_c - \widehat{W}_a \tag{31}$$

قضیه 2: فرض کنید که دینامیک سیستم با رابطه (1) بیان شود. همچنین شبکه‌ی ارزیاب و عملگر بر اساس رابطه (14) عمل کنند و وزن‌های شبکه‌ی عصبی ارزیاب به صورت رابطه (32) به‌هنگام‌سازی شوند.

$$\dot{\widehat{W}}_c = -a\,\frac{\sigma_a}{(\sigma_a^{\mathrm{T}}\sigma_a + 1)^2}\left[\sigma_a^{\mathrm{T}}\widehat{W}_c + Q(x) + \hat{u}^{\mathrm{T}}R\hat{u}\right] \tag{32}$$

که در آن $\sigma_a = \nabla\phi(f + g\hat{u})$ می‌باشد و $\hat{u}$ با رابطه (30) بدست می‌آید. همچنین فرض کنید که شرط تحریک دائمی برای $\bar{\sigma}_a = \frac{\sigma_a}{(\sigma_a^{\mathrm{T}}\sigma_a+1)}$ برقرار است و قانون به‌هنگام‌سازی وزن‌های شبکه‌ی عملگر مطابق رابطه (33) باشد.

$$\begin{aligned}\dot{\widehat{W}}_a &= -b\left\{(F_2\widehat{W}_a - F_1\widehat{W}_c) - \frac{1}{4}\bar{D}_1(x)\widehat{W}_a m^{\mathrm{T}}(x)\widehat{W}_c\right\}\\ \bar{D}_1(x) &\equiv \nabla\phi(x)g(x)R^{-1}g^{\mathrm{T}}(x)\nabla\phi^{\mathrm{T}}(x)\\ m &\equiv \frac{\sigma_a}{(\sigma_a^{\mathrm{T}}\sigma_a + 1)^2}\end{aligned} \tag{33}$$

که در اینجا $F_1 > 0$ و $F_2 > 0$ پارامترهای طراحی هستند، آنگاه N_0 ای وجود دارد که برای تعداد نورون‌های لایه‌ی پنهان بیش از آن $N > N_0$، متغیرهای وضعیت حلقه بسته‌ی سیستم، خطای شبکه‌ی عصبی ارزیاب $\widetilde{W}_c$، و عملگر $\widetilde{W}_a$، محدود نهایی یکنواخت هستند. هم چنین با $\varepsilon_{\max}$ که در اثبات تعریف می‌شود قضیه برقرار می‌باشد؛ به همین دلیل $\widehat{W}_c$ به صورت نمایی به مقادیر بهینه‌ی تقریبی ارزیاب W_c همگرا می‌شوند. (**اثبات**: پیوست الف.)

4-تخمین زدن دینامیک سیستم با استفاده از شناساگر عصبی-مقاوم

روش توضیح داده شده در بخش قبل برای حل مسأله کنترل بهینه، با فرض

[1] Uniform Complete Observability (UCO)
[2] Observability gramian
[3] Residual set

معلوم بودن $f(x)$ و $g(x)$ در معادله (1) بودند که فرضی محدود کننده می‌باشد. در $f(x)$ عباراتی است که تعیین دقیق آن‌ها بسیار مشکل و در مواردی غیر ممکن می‌باشد که از جمله می‌توان به اثرات اصطکاک و میرایی اشاره کرد. از این‌رو در این بخش با استفاده از شناساگر عصبی-مقاوم روشی برای تخمین زدن $f(x)$ ارائه می‌شود و یکی از این فرض‌های محدود کننده برداشته می‌شود و به این ترتیب روش ارائه شده توسط وموداکیس و لوییس [9] بهبود می‌یابد. منظور از شناساگر عصبی- مقاوم آن است که این شناساگر از دو بخش تشکیل شده است؛ بخش شبکه عصبی که وظیفه تقریب زدن دینامیک داخلی را دارد یک شبکه دولایه است که توابع فعالیت آن بصورت سیگموید دو قطبی انتخاب شده‌اند. بخش مقاوم نیز موجب می‌شود شناساگر در مقابل عدم قطعیت‌ها و اغتشاش و همچنین خطای حاصل از تقریب شبکه عصبی مقاوم شود و بتواند شناسایی را به درستی انجام دهد که برای کارهای عملی ساختار مناسبی را ارائه می‌کند. بهنگام شدن وزن‌های شبکه‌ی عصبی و همچنین تغییر مقدار جمله‌ی مقاوم براساس مقدار خطای میان متغیر وضعیت واقعی و تخمینی است. شکل 1 چگونگی استفاده از این شبکه در کنار ساختار عملگر-ارزیاب را نشان می‌دهد.

باتوجه به توضیحات گفته شده در بخش قبل، دینامیک را می‌توان با استفاده از یک شبکه‌ی عصبی مطابق رابطه (34) تقریب زد.

$$\dot{x} = W_f^{\mathrm{T}}\sigma\left(V_f^{\mathrm{T}}x\right) + \varepsilon_f(x) + g(x)\hat{u} \tag{34}$$

که در آن به‌جای $f(x)$ از یک شبکه‌ی عصبی استفاده شده است. در رابطه (34)، $W_f \in \Re^{L_f+1\times n}$، $V_f \in \Re^{n\times L_f}$ وزن‌های ایده‌آل شبکه عصبی هستند که مقدار آن‌ها معلوم نیست. همچنین تابع فعالیت شبکه‌ی عصبی $\sigma_f \triangleq \sigma\left(V_f^{\mathrm{T}}\right) \in \Re^{L_f+1}$ می‌باشد و $\varepsilon_f(x) \in \Re^n$ خطای بازسازی تابع می‌باشد.

برای تقریب رابطه (34) از شبکه‌ی عصبی چندلایه استفاده می‌شود که در رابطه (35) بیان شده است.

$$\dot{\hat{x}} = \widehat{W}_f^{\mathrm{T}}\hat{\sigma}_f + g(x)\hat{u} + \mu \tag{35}$$

در رابطه (35)، $\hat{x}(t) \in \Re^n$ متغیر وضعیت شبکه‌ی عصبی می‌باشد. همچنین $\sigma_f \triangleq \sigma\left(\hat{V}_f^{\mathrm{T}}\hat{x}\right) \in \Re^{L_f+1}$ و $\widehat{W}_f \in \Re^{L_f+1\times n}$ وزن‌های تخمینی شبکه و $\hat{V}_f(t) \in \Re^{n\times L_f}$ می‌باشند. در رابطه (35)، $\mu(t) \in \Re^n$ جمله‌ی پسخور انتگرال مقاوم علامت خطا می‌باشد که در [11] و [12] ارائه شده است. استفاده از این جمله باعث می‌شود که در حضور اغتشاش و عدم قطعیت‌های سیستم نیز بتوان به‌صورت مناسبی دینامیک سیستم را تخمین زد که در رابطه (36) تعریف شده است.

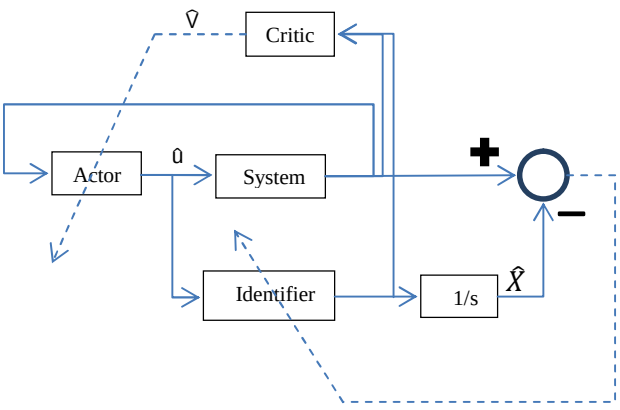


Fig. 1 Schematic of Actor-Critic-Identifier structure

شکل 1 شماتیک ساختار عملگر- ارزیاب - شناساگر

$$\mu(t) \triangleq k\tilde{x}(t) - k\tilde{x}(0) + \upsilon \tag{36}$$

در رابطه (36)، $\tilde{x} \triangleq x(t) - x(t)$ خطای شناسایی می‌باشد و $\upsilon(t) \in \Re^n$ حل معادله دیفرانسیل می‌باشد که در رابطه (37) بیان شده است.

$$\dot{\upsilon} = (k\alpha + \gamma)\tilde{x} + \beta_1\mathrm{sgn}(\tilde{x}); \quad \upsilon(0) = 0 \tag{37}$$

در رابطه (37)، k، α، γ و β_1 اعداد ثابت طراحی هستند و $\mathrm{sgn}(\cdot)$ تابع علامت می‌باشد.

دینامیک خطای شناسایی در رابطه (38) تعریف شده است.

$$\dot{\tilde{x}} = W_f^{\mathrm{T}}\sigma_f - \widehat{W}_f^{\mathrm{T}}\hat{\sigma}_f - \mu \tag{38}$$

خطای شناسایی فیلتر شده مطابق رابطه (39) است.

$$e_f = \dot{\tilde{x}} + \alpha\tilde{x} \tag{39}$$

مشتق زمانی رابطه (39)با استفاده از رابطه (38) برابر می‌شود با رابطه (40).

$$\begin{aligned}\dot{e}_f &= W_f^{\mathrm{T}}\sigma'_f V_f^{\mathrm{T}}\dot{x} - \dot{\widehat{W}}_f^{\mathrm{T}}\hat{\sigma}_f - \widehat{W}_f^{\mathrm{T}}\widehat{\sigma'}_f\dot{\hat{V}}_f^{\mathrm{T}}\hat{x} - \widehat{W}_f^{\mathrm{T}}\widehat{\sigma'}_f\hat{V}_f^{\mathrm{T}}\dot{\hat{x}} + \dot{\varepsilon}_f(x)\\ &\quad -ke_f - \gamma\tilde{x} - \beta_1\mathrm{sgn}(\tilde{x}) + \alpha\dot{\tilde{x}}\end{aligned} \tag{40}$$

براساس رابطه (40) و تحلیل پایداری، وزن‌های شبکه‌ی عصبی شناساگر به صورت رابطه (41) به‌هنگام می‌شوند.

$$\begin{aligned}\dot{\widehat{W}}_f &= \mathrm{proj}(\Gamma_{wf}\widehat{\sigma'}_f\hat{V}_f^{\mathrm{T}}\dot{\hat{x}}\tilde{x}^{\mathrm{T}})\\ \dot{\hat{V}}_f &= \mathrm{proj}(\Gamma_{vf}\dot{\hat{x}}\tilde{x}^{\mathrm{T}}\widehat{W}_f^{\mathrm{T}}\widehat{\sigma'}_f)\end{aligned} \tag{41}$$

در رابطه (41)، $\Gamma_{wf} \in \Re^{(L_f+1)\times(L_f+1)}$ و $\Gamma_{vf} \in \Re^{n\times n}$ ماتریس‌های مثبت ثابت معین می‌باشند [12]. همچنین proj یک عملگر ریاضی تصویر است و هدف استفاده از آن جلوگیری از خارج شدن وزن‌ها از محدوده‌ای است که برای آن‌ها در نظر گرفته شده است. این مسأله زمانی اتفاق می‌افتد که محدوده‌ای برای پارامتری که تخمین زده می‌شود معلوم باشد. برای روشن شدن عملکرد این عملگر فرض می‌شود که برای پارامتر فرضی ζ حد بالا و پایین مشخص باشد، به عبارت دیگر $\zeta \in [p^-, p^+]$، در اینصورت عملکرد این عملگر مطابق رابطه (42) است [16].

$$\dot{\zeta} = 0 \quad if \begin{cases}\zeta = p^- & and\ \dot{\zeta} < 0\\ \zeta = p^+ & and\ \dot{\zeta} > 0\end{cases} \tag{42}$$

5-نتایج

در این بخش دو مثال مورد بررسی قرار می‌گیرد. در مثال اول، به صورت تجربی کنترل موتور DC به عنوان یک سیستم خطی مورد بررسی قرار می‌گیرد و در مثال دوم شبیه‌سازی یک سیستم غیرخطی ارائه می‌شود.

5-1-سیستم خطی

موتور DC و بخش‌های مختلف آن در شکل 2 نشان داده شده است.

برای این موتور دو سناریو مورد بررسی قرار گرفت. در ابتدا برای موتور با دیسک سبک، شناسایی پارامتر انجام شد. به این ترتیب پارامترهای معادلات دینامیکی موتوربدست آمد. نتایج در رابطه (43) بیان شده است.

$$\begin{cases}\dot{\theta} = \omega\\ \dot{\omega} = -\dfrac{B}{J}\omega + \dfrac{K_t}{J}i\\ \dot{i} = -\dfrac{K_e}{L_a}\omega - \dfrac{R_a}{L_a}i + \dfrac{1}{L_a}v\end{cases}$$

$$\frac{B}{J} = 5.3542, \frac{K_t}{J} = 28.0626, \frac{K_e}{L_a} = 18.6812, \frac{R_a}{L_a} = 47.516,$$
$$\frac{1}{L_a} = 112.3207 \tag{43}$$

Fig. 2 DC motor

شکل 2 موتور DC

در رابطه (43)، θ موقعیت، ω سرعت، i جریان موتور و v ولتاژ ورودی می‌باشد. پارامترهای سیستم نیز به این صورت تعریف شده‌اند؛ B ضریب میرایی، J ممان اینرسی، K_t ثابت گشتاور، K_e ثابت سرعت، R_a مقاومت و L_a القایی موتور می‌باشد.

بر اساس این مقادیر، با حل معادله ریکاتی کنترلر بهینه برای سیستم طراحی شد. همچنین با استفاده از ساختار عملگر-ارزیاب-شناساگر و به روش برخط مسأله کنترل بهینه برای این سیستم حل شد. انتظار می‌رود که در صورت آنکه پارامترهای سیستم به درستی شناسایی شده باشند، نتایج حاصل از این دو کنترلر با یکدیگر تطابق خوبی داشته باشند.

ماتریس‌های وزنی تابع هزینه به صورت رابطه (44) انتخاب شده است.

$$Q = \mathrm{diag}\{20,0.5,0.1\},\quad R = 1 \tag{44}$$

با حل معادله ریکاتی، ضرایب بهره‌ی بهینه بدست آمدند که در رابطه (45) بیان شده است.

$$k = \{4.4721,0.5954,0.3362\} \tag{45}$$

برای حل برخط مسأله با استفاده از ساختار عملگر-ارزیاب-شناساگر، وزن‌های اولیه، پارامترهای طراحی و توابع پایه‌ی شبکه‌های عصبی مطابق رابطه (46) انتخاب شدند. توجه شود که وزن‌های اولیه به صورتی انتخاب شدند که پایدار کننده باشند.

$$\begin{aligned} &a = 10,\quad b = 3,\quad F_1 = F_2 = 25 \\ &\phi(x) = [x_1^2, x_1x_2, x_1x_3, x_2x_3, x_2^2, x_3^2] \end{aligned} \tag{46}$$

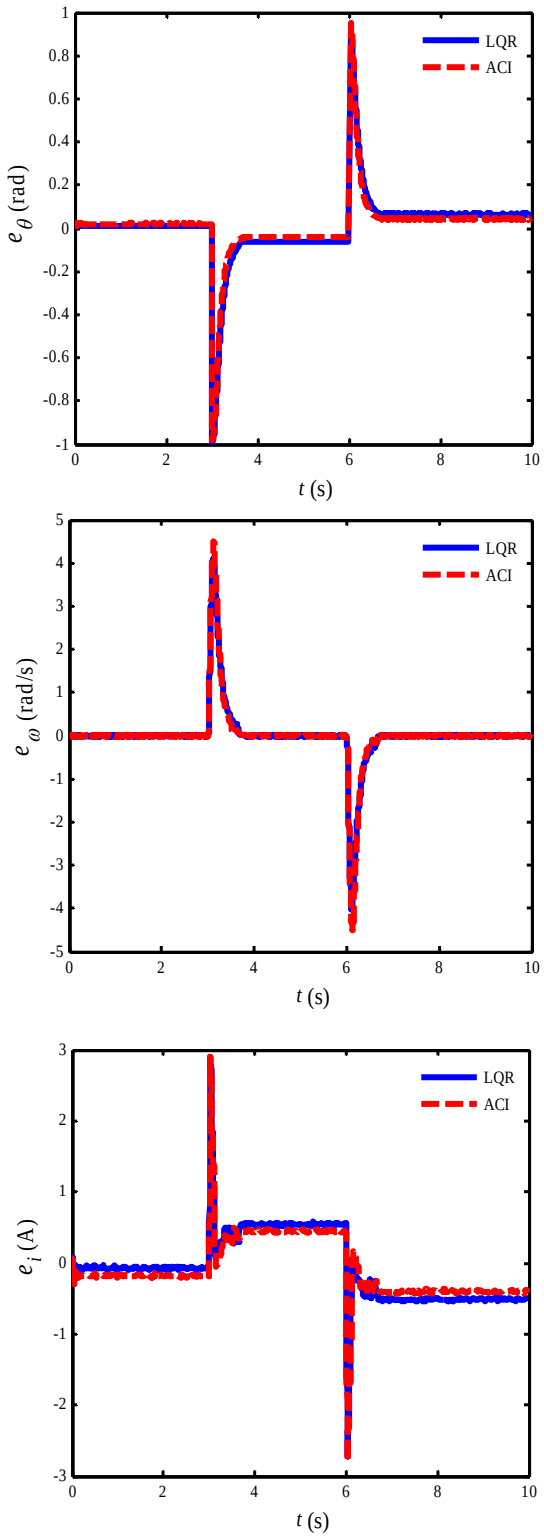
به منظور برقراری شرط تحریک دائمی، به مدت 100 ثانیه سیگنال بیان شده در رابطه (47) که فرکانس‌های مختلف را در بر دارد و برحسب زمان از اثر آن کاسته می‌شود به ورودی سیستم افزوده شد.

$$\begin{aligned} n = {} & 4\exp(-0.009t)(0.3\sin(8t)^2\cos(2t) \\ & +\sin(-1.2t)^2\cos(0.5t)) \end{aligned} \tag{47}$$

وجود فرکانس‌های مختلف موجب تحریک بخش‌های مختلف سیستم و با دامنه‌های تحریک گوناگون می‌شود که به غنای مسیر جهت برقراری شرط تحریک دائمی کمک می‌کند. همچنین مدت زمان اعمال آن (در اینجا 100 ثانیه) با توجه به سعی و خطا و مشاهده نحوه همگرایی وزن‌های شبکه عصبی انتخاب شده است.

خطای حاصل از این دو کنترلر در تعقیب یک ورودی پالسی در شکل 3 نشان داده شده است.

در مرحله‌ی بعد با تغییر دیسک و استفاده از دیسک سنگین، در پارامتر سیستم (ممان اینرسی) تغییر ایجاد کرده و با استفاده از دو کنترلر برون خط (LQR) و برخط (ACI) به کنترل سیستم برای تعقیب ورودی پالسی پرداخته می‌شود. انتظار می‌رود که کنترلر برخط باتوجه به اینکه حین دوره‌ی یادگیری دینامیک سیستم را یاد می‌گیرد نسبت به کنترلر برون خط که برای دیسک سبک (ممان اینرسی کمتر نسبت به دیسک سنگین) طراحی شده نتایج مناسب‌تری بدست دهد.



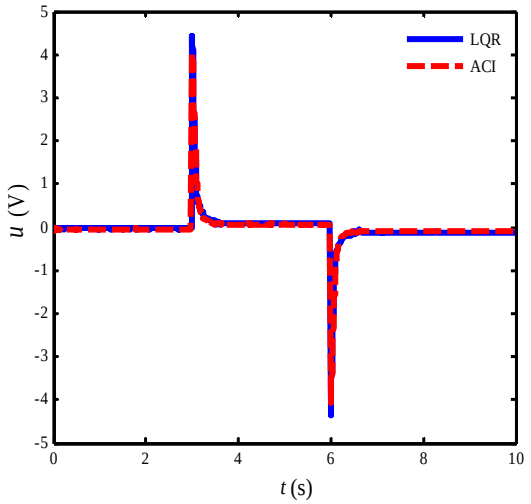
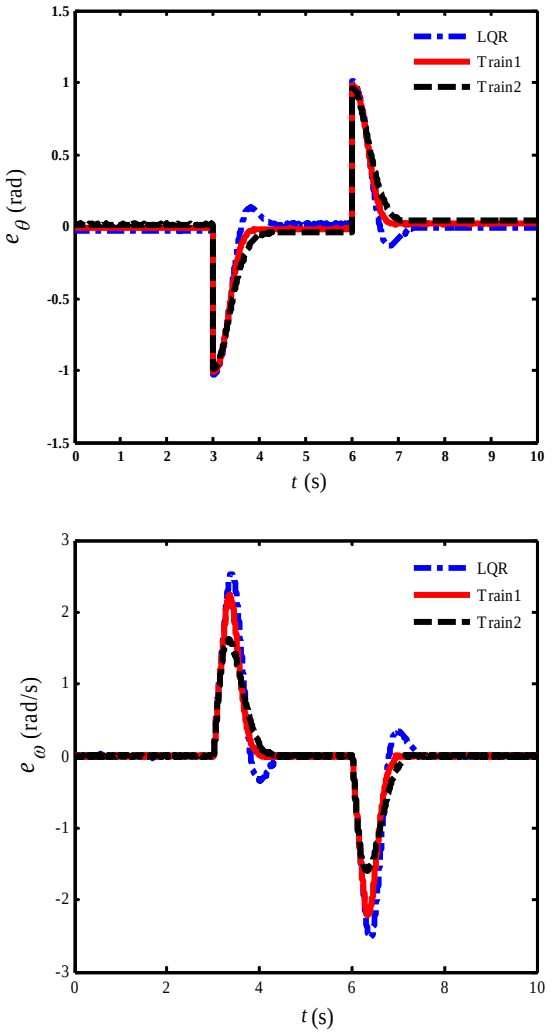


Fig. 3 Comparison between ACI and LQR controllers – light disk

شکل 3 مقایسه‌ی میان کنترلرهای ACI و LQR – دیسک سبک

در این بخش نیز یادگیری در دو دوره و هرکدام به مدت 100 ثانیه با استفاده از سیگنال نویزی که برای دیسک سبک بکار رفت انجام شده است. دیگر پارامترهای شبکه‌ی عصبی نیز مانند حالت اول است (شکل 4).



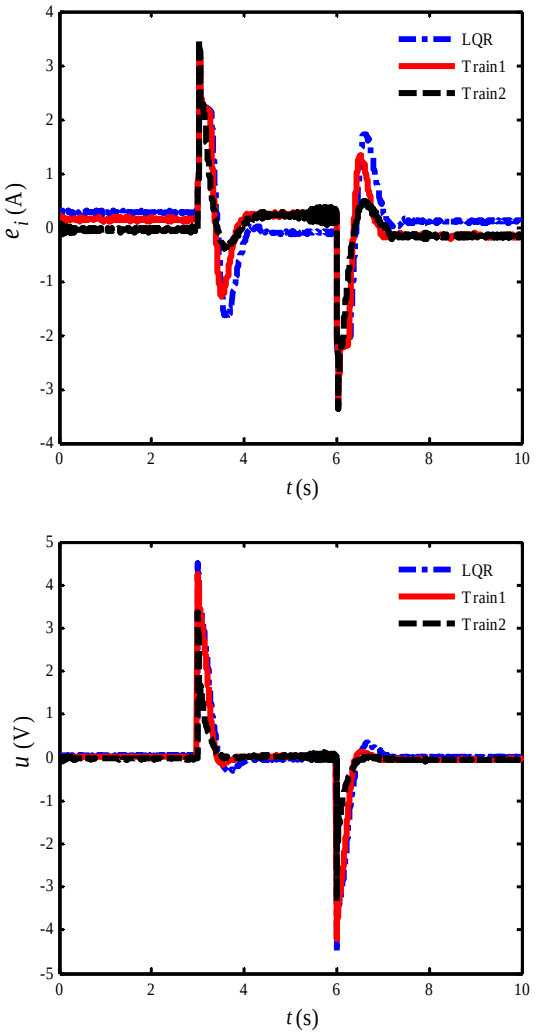


Fig. 4 Comparison between ACI and LQR controllers – heavy disk

شکل 4 مقایسه‌ی میان کنترلرهای ACI و LQR – دیسک سنگین

مقایسه‌ی میان عملکرد کنترلرهای برخط پس از یادگیری اول و دوم قابل توجه است. هرچند که در نمودار مربوط به خطای موقعیت، خطای کنترلر اول کمتر از کنترلر دوم است اما با توجه به شکل‌های سرعت، جریان و ورودی کنترلی مشخص می‌شود که عملکرد کنترلر دوم بهتر از اولی می‌باشد. برای مقایسه‌ی مناسبتر عملکرد این سه کنترلر هزینه‌ی آن‌ها در این تعقیب محاسبه شد و نتیجه به‌صورت رابطه (48) می‌باشد.

هزینه:

$$\text{LQR} = 19.8317, \text{Train1} = 17.821, \text{Train2} = 14.9216 \tag{48}$$

حال برای یافتن علت بهتر بودن پاسخ‌های کنترلر برخط، ضرایب بهره‌ی کنترلی مورد بررسی قرار می‌گیرد. این ضرایب، برای کنترلر برخط با استفاده از رابطه (14) بدست می‌آید. کنترلرهای بهینه که از حل معادله ریکاتی و حل برخط بدست آمده‌اند در رابطه (49) آورده شده‌اند.

$$\begin{aligned} u_{\text{LQR}} &= -(4.4721x_1 + 0.5954x_2 + 0.3362x_3) \\ u_{\text{Train1}} &= -(4.2289x_1 + 0.9154x_2 + 0.1685x_3) \\ u_{\text{Train2}} &= -(3.4129x_1 + 1.0743x_2 + 0.5256x_3) \end{aligned} \tag{49}$$

همان‌طورکه در شکل 4 مشاهده می‌شود با تغییر دیسک، پاسخ کنترلر برون خط با فراجهش همراه بود. طبیعی است که برای کاهش فراجهش باید ضریب کنترلی برای سرعت را افزایش داد (مشابه افزایش ترم مشتقی در کنترلر

تناسبی- مشتقی). مشاهده می‌شود که پس از دوره‌ی اول یادگیری این ضریب افزایش یافته است و این مسأله منجر به از بین رفتن فراجهش شده است. زمانیکه فراجهش از بین می‌رود نیازی به اعمال ورودی بالا برای تعقیب وجود ندارد. در کنترلر برون خط بخشی از این ورودی بالا صرف فراجهش شده است. طبیعی است که در یادگیری اول با از بین رفتن فراجهش، این نیاز به ورودی بالا نیز از بین رفته و ضرایب کنترلی موقعیت و جریان نیز کاهش یابد.

با مقایسه میان مرحله‌ی اول و دوم یادگیری مشخص است که ضریب کنترلی سرعت و جریان افزایش یافته در حالیکه ضریب مکان کاهش یافته است. این مسأله را می‌توان به اینصورت توضیح داد که باتوجه به این موضوع که هدف کاهش هزینه است، کنترلر تلاش کرده تا از میزان سرعت و جریان بکاهد. همچنین به دلیل آنکه سرعت بالای صفر شدن خطای موقعیت موجب بالا رفتن ورودی کنترلی نشود، با کاهش ضریب کنترلی مکان اجازه داده شده که این خطا نسبت به یادگیری اول دیرتر صفر شود و در ازای آن هزینه بیشتر کاهش یابد.

2-5-سیستم غیرخطی

در این بخش روش پیشنهادی برای یک مسأله کنترل بهینه که دارای حل دقیق و تحلیلی است پیاده‌سازی می‌شود. سیستم غیرخطی بیان شده در رابطه (50) در نظر گرفته می‌شود [9].

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1-(\cos(2x_1)+2)^2) \end{bmatrix} + \begin{bmatrix} 0 \\ \cos(2x_1)+2 \end{bmatrix} u \quad (50)$$

ماتریس‌های وزنی در تابع هزینه مطابق رابطه (51) درنظر گرفته شده‌اند.

$$Q = \mathrm{diag}\{1,1\}, R = 1 \quad (51)$$

تابع ارزش و کنترل بهینه‌ی دقیق و تحلیلی این مسأله در رابطه (52) آورده شده‌اند.

$$V^*(x) = \frac{1}{2}x_1^2 + x_2^2, u^*(x) = -(\cos(2x_1)+2)x_2 \quad (52)$$

برای استفاده از قضیه ، توابع فعالیت شبکه‌ی عصبی و پارامترهای طراحی به صورت رابطه (53) در نظر گرفته شده است.

$$\phi(x) = x_1^2, x_1x_2, x_2^2 \quad a = 25, b = 10, F_1 = F_2 = 15 \quad (53)$$

بنابراین در حل دقیق مسأله ضرایب ارزیاب برابر رابطه (54) است.

$$W_c = \begin{bmatrix} 0.5 & 0 & 1 \end{bmatrix} \quad (54)$$

پارامترهای طراحی برای شبکه‌ی شناساگر نیز به صورت رابطه (55) انتخاب شده است.

$$k = 80, \alpha = 300, \gamma = 5, \beta = 0.2\Gamma_{wf} = 0.1\mathrm{I}_{6\times6}, = 0.1\mathrm{I}_{4\times4} \quad (55)$$

این مقادیر با توجه به [13] و سعی و خطا انتخاب شده‌اند. همچنین توابع فعالیت برای نورون‌های لایه‌ی پنهان، سیگموید دو قطبی درنظر گرفته شده است که کاربرد وسیعی در شبکه‌ی عصبی و کنترل دارد [14].

همان‌طور که در بخش‌های قبل بیان شد برای همگرایی وزن‌های شبکه‌های عصبی نیاز است که شرط تحریک دائمی برقرار شود. از آنجاکه در سیستم‌های غیرخطی روش مشخصی برای چک کردن برقراری تحریک دائمی نیست، سیگنالی نویزی که از ورودی‌های سینوس و کسینوس با فرکانس‌های مختلف تشکیل شده است برای چند ثانیه با جمع بر ورودی کنترلی به ورودی سیستم اعمال می‌شود. در این مسأله نیز به مدت 20 ثانیه سیگنالی مطابق رابطه (56) به ورودی سیستم اعمال می‌شود.

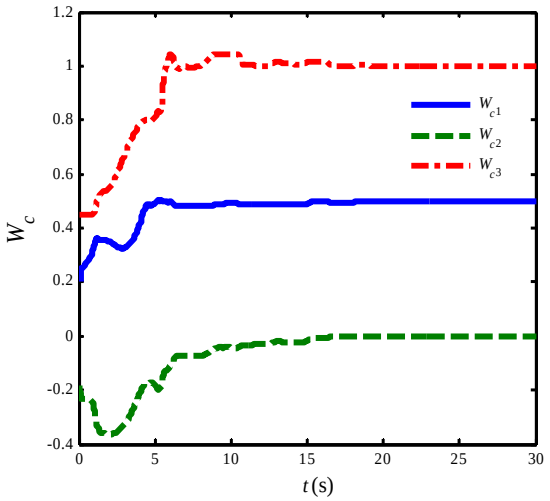


Fig. 5 Convergence of the critic weights to analytic solution

شکل 5 یادگیری شبکه‌ی ارزیاب و همگرایی ضرایب به حل دقیق تحلیلی

$$n(t) = 5\exp(-0.009t)(\sin(t)^2\cos(t) + \sin(2t)^2\cos(0.1t) + \sin(-1.2t)^2\cos(0.5t) + \sin(t)^5) \quad (56)$$

مقادیر اولیه‌ی وزن‌ها به صورت تصادفی در بازه‌ی [1و1-] انتخاب شده‌اند. همگرایی وزن‌های شبکه‌ی ارزیاب در شکل 5 مشاهده می‌شود.

همچنین متغیرهای وضعیت سیستم و تلاش کنترلی حین یادگیری در شکل 6 ارائه شده‌اند.

مشاهده می‌شود که وزن‌ها به مقادیر بهینه‌ی خود همگرا شده‌اند. مقادیر نهایی وزن‌ها در رابطه (57) بیان شده است.

$$W = [0.5002 \quad -0.001 \quad 1.0005] \quad (57)$$

با استفاده از این وزن‌ها کنترل بهینه با رابطه (58) برابر می‌شود.

$$u = -\frac{1}{2}R^{-1}\begin{bmatrix} 0 \\ \cos(2x_1)+2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2x_1 & 0 \\ x_2 & x_1 \\ 0 & 2x_2 \end{bmatrix} \begin{bmatrix} 0.5002 \\ -0.0010 \\ 1.0005 \end{bmatrix} \quad (58)$$

که به مقدار بهینه‌ی خود بسیار نزدیک است. همچنین تابع هزینه‌ی بهینه نهایی بدست آمده به صورت رابطه (59) محاسبه می‌شود.

$$V = [0.5002 \quad -0.0010 \quad 1.0005] \begin{bmatrix} x_1^2 \\ x_1x_2 \\ x_2^2 \end{bmatrix} \quad (59)$$

خطای متغیرهای وضعیت سیستم و شناساگر در شکل 7 نشان داده شده است. همان‌طور که از شکل مشخص است شناساگر خیلی سریع توانسته است که دینامیک سیستم را یاد بگیرد.

6-بحث و نتیجه‌گیری

در این مقاله روشی برای حل معادله همیلتون-ژاکوبی-بلمن به صورت برخط و بدون نیاز به دانستن دینامیک داخلی سیستم ارائه شده است. حل تقریبی معادله همیلتون-ژاکوبی-بلمن با استفاده از دو شبکه‌ی عصبی در ساختار عملگر-ارزیاب که به ترتیب قانون کنترلی بهینه و تابع ارزش بهینه را تقریب می‌زنند صورت گرفته است. به‌منظور تخمین دینامیک داخلی سیستم از یک شبکه‌ی عصبی-مقاوم استفاده شده است. شبیه‌سازی‌ها برای یک سیستم غیرخطی که حل دقیق همیلتون-ژاکوبی-بلمن آن معلوم است، کارایی این روش را به وضوح نشان می‌دهد. همچنین با پیاده‌سازی این کنترلر به صورت عملی برای موتور DC، مزیت این کنترلر نسبت به کنترلر برون خط نشان داده شد. در ادامه می‌توان این روش را به حالتی تعمیم داد که اشباع

عملگرها را نیز در تابع هزینه درنظر گرفته و در حضور اشباع، بهینه‌سازی بررسی شود.

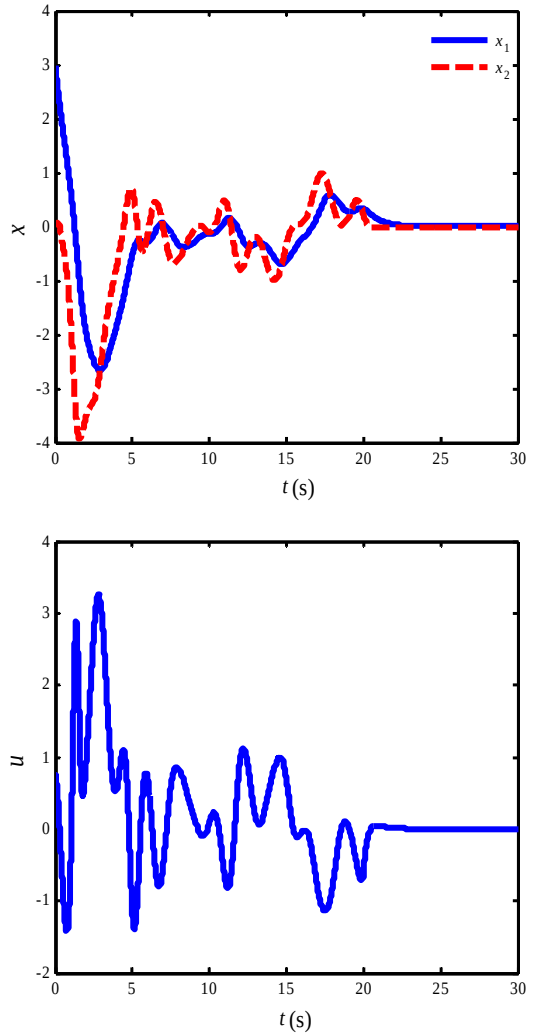


Fig. 6 States and control input during learning period

شکل 6 متغیرهای وضعیت و تلاش کنترلی حین یادگیری

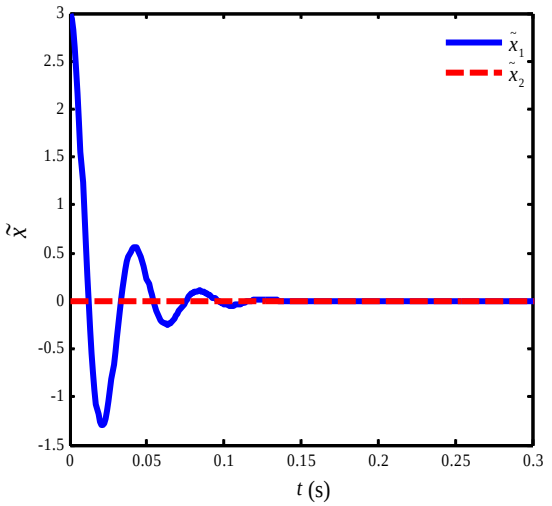


Fig. 7 Error between actual and estimated states

شکل 7 خطای متغیرهای وضعیت سیستم و شناساگر

7-پیوست الف: اثبات قضیه 2

اثبات پیش‌رو برگرفته از [9] می‌باشد که به دلیل تغییر پارامترها در این مقاله، برخی اشکالات و خلاصه نویسی اثبات‌ها در مرجع اصلی بازنویسی شده است.

به منظور اثبات این قضیه، تابع لیاپانوفی مطابق رابطه (الف- 1) انتخاب می‌شود.

$$L(x) = V(x) + \tfrac{1}{2}tr(\widetilde{W}_c^{\mathrm{T}} a^{-1}\widetilde{W}_c) + \tfrac{1}{2}tr(\widetilde{W}_a^{\mathrm{T}} b^{-1}\widetilde{W}_a). \qquad \text{(الف- 1)}$$

مشتق تابع (الف- 1) به صورت رابطه (الف- 2) می‌باشد.

$$\dot{L}(x) = \dot{V}(x) + \widetilde{W}_c^{\mathrm{T}} a^{-1}\dot{\widetilde{W}}_c + \widetilde{W}_a^{\mathrm{T}} b^{-1}\dot{\widetilde{W}}_a = \dot{L}_V(x) + \dot{L}_1(x) + \dot{L}_2(x) \qquad \text{(الف- 2)}$$

جمله‌ی اول رابطه (الف- 2) برابر است با رابطه (الف- 3).

$$\dot{V}(x) = W_c^{\mathrm{T}}\left(\nabla\phi_1 f(x) - \frac{1}{2}\bar{D}_1(x)\widehat{W}_a\right) + \nabla\varepsilon^{\mathrm{T}}(x)\left(f(x) - \tfrac{1}{2}g(x)R^{-1}g^{\mathrm{T}}(x)\nabla\phi_1^{\mathrm{T}}\widehat{W}_a\right) \qquad \text{(الف- 3)}$$

بنابراین:

$$\begin{aligned}\dot{V}(x) &= W_c^{\mathrm{T}}\left(\nabla\phi_1 f(x) - \frac{1}{2}\bar{D}_1(x)\widehat{W}_a\right) + \varepsilon_1(x) = W_c^{\mathrm{T}}\nabla\phi_1 f(x) \\ &+ \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(x)(W_c - \widehat{W}_a) - \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \varepsilon_1(x) \\ &= W_c^{\mathrm{T}}\nabla\phi_1 f(\mathrm{x}) + \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(\mathrm{x})\widetilde{W}_a - \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(\mathrm{x})W_c + \varepsilon_1(\mathrm{x}) \\ &= W_c^{\mathrm{T}}\sigma_1 + \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(\mathrm{x})\widetilde{W}_a + \varepsilon_1(\mathrm{x})\end{aligned} \qquad \text{(الف- 4)}$$

که در آن:

$$\varepsilon_1(x) \equiv \dot{\varepsilon}(x) = \nabla\varepsilon^{\mathrm{T}}(x)f(x) - \tfrac{1}{2}g(x)R^{-1}g^{\mathrm{T}}(x)\nabla\phi_1^{\mathrm{T}}\widehat{W}_a \qquad \text{(الف- 5)}$$

از معادله همیلتون-ژاکوبی-بلمن، رابطه (15)، رابطه (الف- 6) بدست می‌آید.

$$W_c^{\mathrm{T}}\sigma_1 = -Q(x) - \frac{1}{4}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \varepsilon_{\mathrm{HJB}}(x) \qquad \text{(الف- 6)}$$

بنابراین جمله‌ی اول رابطه (الف- 2) برابر می شود با رابطه (الف- 7).

$$\dot{L}_V(x) = -Q(x) - \frac{1}{4}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \varepsilon_{\mathrm{HJB}}(x) + \varepsilon_1(x) + \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(x)\widetilde{W}_a \equiv \dot{\bar{L}}_V(x) + \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(x)\widetilde{W}_a + \varepsilon_1(x) \qquad \text{(الف- 7)}$$

جمله‌ی دوم رابطه (الف- 2) با محاسباتی که در ادامه می‌آید به صورت رابطه (الف- 8) قابل بیان است.

$$\begin{aligned}\dot{L}_1(x) &= \widetilde{W}_c^{\mathrm{T}} a^{-1}\dot{\widetilde{W}}_c = \widetilde{W}_c^{\mathrm{T}} a^{-1} a\,\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2 + 1)^2}\left(\sigma_2^{\mathrm{T}}\widehat{W}_c + Q(x) + \frac{1}{4}\widehat{W}_a^{\mathrm{T}}\bar{D}_1\widehat{W}_a\right) \\ &= \widetilde{W}_c^{\mathrm{T}}\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2 + 1)^2}\left(\sigma_2^{\mathrm{T}}\widehat{W}_c + Q(x) + \frac{1}{4}\widehat{W}_a^{\mathrm{T}}\bar{D}_1\widehat{W}_a - Q(x) - \sigma_1^{\mathrm{T}}W_c - \frac{1}{4}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \varepsilon_{\mathrm{HJB}}(x)\right) \\ &= \widetilde{W}_c^{\mathrm{T}}\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2 + 1)^2}\left(\sigma_2^{\mathrm{T}}\widehat{W}_c - \sigma_1^{\mathrm{T}}W_c + \frac{1}{4}\widehat{W}_a^{\mathrm{T}}\bar{D}_1\widehat{W}_a - \frac{1}{4}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \varepsilon_{\mathrm{HJB}}(x)\right) \\ &= \widetilde{W}_c^{\mathrm{T}}\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2 + 1)^2}\left(-\widetilde{W}_c^{\mathrm{T}}(\nabla\phi_1^{\mathrm{T}})f(x) - \frac{1}{2}\widehat{W}_a^{\mathrm{T}}\bar{D}_1\widehat{W}_c + \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \frac{1}{4}\widehat{W}_a^{\mathrm{T}}\bar{D}_1\widehat{W}_a - \frac{1}{4}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \varepsilon_{\mathrm{HJB}}(x)\right) \\ &= \widetilde{W}_c^{\mathrm{T}}\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2+1)^2}\left(-f(x)^{\mathrm{T}}\nabla\phi_1^{\mathrm{T}}\widetilde{W}_c + \tfrac{1}{2}\widehat{W}_a^{\mathrm{T}}\bar{D}_1\widehat{W}_c + \tfrac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1\widetilde{W}_a + \varepsilon_{\mathrm{HJB}}(x)\right)\end{aligned}$$

$$\dot{L}_1(x) = \widetilde{W}_c^{\mathrm{T}} \frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2+1)^2}\left(-\sigma_2^{\mathrm{T}}\widetilde{W}_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1\widetilde{W}_a + \varepsilon_{\mathrm{HJB}}(x)\right)$$
$$= \dot{\bar{L}}_1 + \frac{1}{4}\widetilde{W}_c^{\mathrm{T}}\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2+1)^2}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1\widetilde{W}_a \qquad \text{(الف- 8)}$$

که در آن:

$$\dot{\bar{L}}_1 = \widetilde{W}_c^{\mathrm{T}}\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2+1)^2}\left(-\sigma_2^{\mathrm{T}}\widetilde{W}_c + \varepsilon_{\mathrm{HJB}}(x)\right) = \widetilde{W}_c^{\mathrm{T}}\bar{\sigma}_2\left(-\bar{\sigma}_2^{\mathrm{T}}\widetilde{W}_c + \frac{\varepsilon_{\mathrm{HJB}}(x)}{m_s}\right)$$

سرانجام با اضافه کردن روابط (الف- 7) و (الف- 8)، رابطه (الف- 9) بدست می‌آید.

$$\dot{L}(x) = -Q(x) - \frac{1}{4}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \frac{1}{2}W_c^{\mathrm{T}}\bar{D}_1(x)\widetilde{W}_a + \varepsilon_{\mathrm{HJB}}(x) + \varepsilon_1(x) + \widetilde{W}_a^{\mathrm{T}}b^{-1}\dot{\widetilde{W}}_a + \widetilde{W}_c^{\mathrm{T}}\frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2+1)^2}\left(-\sigma_2^{\mathrm{T}}\widetilde{W}_c + \varepsilon_{\mathrm{HJB}}(x) + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1\widetilde{W}_a\right)$$
$$\dot{L}(x) = \dot{\bar{L}}_V + \dot{\bar{L}}_1 + \varepsilon_1(x) - \widetilde{W}_a^{\mathrm{T}}b^{-1}\dot{\widehat{W}}_a + \frac{1}{2}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}\widetilde{W}_c - \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}W_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)\widetilde{W}_a\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}W_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)\widehat{W}_a\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}\widetilde{W}_c \qquad \text{(الف- 9)}$$

که در آن:

$$\bar{\sigma}_2 = \frac{\sigma_2}{(\sigma_2^{\mathrm{T}}\sigma_2+1)^2}, \quad m_s = \sigma_2^{\mathrm{T}}\sigma_2 + 1$$

به منظور انتخاب قانون بهنگام‌سازی شبکه‌ی عصبی عملگر، رابطه (الف-9) به صورت رابطه (الف- 10) بازنویسی می‌شود.

$$\dot{L}(x) = \dot{\bar{L}}_V + \dot{\bar{L}}_1 + \varepsilon_1(x) - \widetilde{W}_a^{\mathrm{T}}\left[b^{-1}\dot{\widehat{W}}_a - \frac{1}{4}\bar{D}_1(x)\widehat{W}_a\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}\widehat{W}_c\right] + \frac{1}{2}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}\widetilde{W}_c - \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}W_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2}{m_s}\widetilde{W}_a \qquad \text{(الف- 10)}$$

قانون بهنگام‌سازی عملگر مطابق رابطه (الف- 11) انتخاب می‌شود.

$$\dot{\widehat{W}}_a = -b\left\{(F_2\widehat{W}_a - F_1\widehat{W}_c) - \frac{1}{4}\bar{D}_1(x)\widehat{W}_a m^{\mathrm{T}}(x)\widehat{W}_c\right\} \qquad \text{(الف- 11)}$$

با این انتخاب به L جملات بیان شده در رابطه (الف- 12) اضافه می‌شود.

$$\widetilde{W}_a^{\mathrm{T}}F_2\widehat{W}_a - \widetilde{W}_a^{\mathrm{T}}F_1\widehat{W}_c = \widetilde{W}_a^{\mathrm{T}}F_2(W_c - \widetilde{W}_a) - \widetilde{W}_a^{\mathrm{T}}F_1(W_c - \widetilde{W}_c) = \widetilde{W}_a^{\mathrm{T}}F_2W_c - \widetilde{W}_a^{\mathrm{T}}F_2\widetilde{W}_a - \widetilde{W}_a^{\mathrm{T}}F_1W_c + \widetilde{W}_a^{\mathrm{T}}F_1\widetilde{W}_c \qquad \text{(الف- 12)}$$

به این ترتیب با جمع کلیه‌ی این عبارات رابطه (الف- 13) بدست می‌آید.

$$\dot{L}(x) = -Q(x) - \frac{1}{4}W_c^{\mathrm{T}}\bar{D}_1(x)W_c + \varepsilon_{\mathrm{HJB}}(x) + \widetilde{W}_c^{\mathrm{T}}\bar{\sigma}_2\left(-\bar{\sigma}_2^{\mathrm{T}}\widetilde{W}_c + \frac{\varepsilon_{\mathrm{HJB}}(x)}{m_s}\right) + \varepsilon_1(x) + \frac{1}{2}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}\widetilde{W}_c - \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2^{\mathrm{T}}}{m_s}W_c + \frac{1}{4}\widetilde{W}_a^{\mathrm{T}}\bar{D}_1(x)W_c\frac{\bar{\sigma}_2}{m_s}\widetilde{W}_a + \widetilde{W}_a^{\mathrm{T}}F_2W_c - \widetilde{W}_a^{\mathrm{T}}F_2\widetilde{W}_a - \widetilde{W}_a^{\mathrm{T}}F_1W_c + \widetilde{W}_a^{\mathrm{T}}F_1\widetilde{W}_c \qquad \text{(الف- 13)}$$

اکنون محدوده‌ی نرم‌ها معرفی می‌شود. با استفاده از فرضیات مطرح شده می‌توان رابطه (الف- 14)را نوشت.

$$\|\varepsilon_1(x)\| < b_{\varepsilon x}b_f\|x\| + \frac{1}{2}b_{\varepsilon x}b_g^2b_{\phi x}\sigma_{\min}(R)(\|W_c\| + \|\widetilde{W}_2\|) \qquad \text{(الف- 14)}$$

هم چنین باتوجه به اینکه $Q(x) > 0$، q ای وجود دارد که برای تمامی $x \in \Omega$ رابطه $x^{\mathrm{T}}qx < Q(x)$ برقرار است. در [5] نشان داده شده است که با افزایش N، $\varepsilon_{\mathrm{HJB}}$ به سمت صفر میل می‌کند.

با انتخاب $\varepsilon > 0$ و $N_0(\varepsilon)$ به‌گونه‌ای که $\sup_{x\in\Omega}\|\varepsilon_{\mathrm{HJB}}\| < \varepsilon$ و با فرض اینکه $N > N_0$ و با $\tilde{Z} = [x \quad \widetilde{W}_c \quad \widetilde{W}_a]^{\mathrm{T}}$ رابطه (الف-13) به رابطه (الف-15) تبدیل می‌شود.

$$\dot{L} < \frac{1}{4}\|W_c\|^2\|\bar{D}_1(x)\| + \varepsilon + \frac{1}{2}\|W_c\|b_{\varepsilon x}b_g^2b_{\phi x}\sigma_{\min}(R) - \tilde{Z}^{\mathrm{T}}M\tilde{Z} + \tilde{Z}^{\mathrm{T}}d \qquad \text{(الف- 15)}$$

که در آن:

$$M = \begin{bmatrix} qI & 0 & 0 \\ 0 & I & \left(-\frac{1}{2}F_1 - \frac{1}{8m_s}\bar{D}_1W_c\right)^{\mathrm{T}} \\ 0 & \left(-\frac{1}{2}F_1 - \frac{1}{8m_s}\bar{D}_1W_c\right) & F_2 - \frac{1}{8}(\bar{D}_1W_cm^{\mathrm{T}} + mW_c^{\mathrm{T}}\bar{D}_1) \end{bmatrix},$$

$$d = \begin{bmatrix} b_{\varepsilon x}b_f \\ \frac{\varepsilon}{m_s} \\ \left(\frac{1}{2}\bar{D}_1 + F_2 - F_1\bar{\sigma}_2^{\mathrm{T}} - \frac{1}{4}\bar{D}_1W_cm^{\mathrm{T}}\right)W_c + \frac{1}{2}b_{\varepsilon x}b_g^2b_{\phi x}\sigma_{\min}(R) \end{bmatrix}$$

پارامتر c مطابق رابطه (الف- 16) تعریف می‌شود.

$$c = \frac{1}{4}\|W_c\|^2\|\bar{D}_1(x)\| + \varepsilon + \frac{1}{2}\|W_c\|b_{\varepsilon x}b_g^2b_{\phi x}\sigma_{\min}(R) \qquad \text{(الف- 16)}$$

فرض شود پارامترها بگونه‌ای انتخاب شوند که $M > 0$. دراینصورت رابطه (الف-15) به صورت رابطه (الف- 17) قابل بیان است.

$$\dot{L} < -\|\tilde{Z}\|^2\sigma_{\min}(M) + \|d\|\|\tilde{Z}\| + c + \varepsilon \qquad \text{(الف- 17)}$$

با مربع کامل کردن رابطه (الف- 17) نتیجه گرفته می‌شود که مشتق تابع لیاپانوف منفی است در صورتی که نامساوی (الف- 18) برقرار باشد.

$$\|\tilde{Z}\| > \frac{\|d\|}{2\sigma_{\min}(M)} + \sqrt{\frac{\|d\|^2}{4\sigma_{\min}^2(M)} + \frac{c+\varepsilon}{\sigma_{\min}(M)}} \equiv B_Z \qquad \text{(الف- 18)}$$

به این ترتیب با استفاده از بسط قضیه‌ی لیاپانوف نتیجه‌گرفته می‌شود که متغیرهای وضعیت و وزن‌ها محدود نهایی یکنواخت هستند [14].

8-مراجع

[1] F. L. Lewis, D. Vrabie, V. L. Syrmos, *Optimal control*: John Wiley & Sons, 2012.

[2] D. Kleinman, On an iterative technique for Riccati equation computations, *IEEE Transactions on Automatic Control*, Vol. 13, No. 1, pp. 114–115, 1968.

[3] G. N. Saridis, C.–S. G. Lee, An Approximation Theory of Optimal Control for Trainable Manipulators, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 9, No. 3, pp. 152–159, 1979.

[4] R. W. Beard, G. N. Saridis, J. T. Wen, Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation, *Automatica*, Vol. 33, No. 12, pp. 2159–2177, 1997.

[5] M. Abu–Khalaf, F. L. Lewis, Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach, *Automatica*, Vol. 41, No. 5, pp. 779–791, 2005.

[6] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*: Bradford Book, 1998.

[7] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, C. Melhuish, Reinforcement learning and optimal adaptive control: An overview and implementation examples, *Annual Reviews in Control*, Vol. 36, No. 1, pp. 42–59, 2012.

[8] D. Vrabie, F. Lewis, Neural network approach to continuous–time direct adaptive optimal control for partially unknown nonlinear systems, *Neural Networks*, Vol. 22, No. 3, pp. 237–246, 2009.

[9] K. G. Vamvoudakis, F. L. Lewis, Online actor–critic algorithm to solve the continuous–time infinite horizon optimal control problem, *Automatica*, Vol. 46, No. 5, pp. 878–888, 2010.

[10] L. Busoniu, R. Babuska, B. De Schutter, D. Ernst, *Reinforcement learning and dynamic programming using function approximators*: CRC Press, 2010.
[11] B. Xian, D. M. Dawson, M. S. De Queiroz, J. Chen, A continuous asymptotic tracking control strategy for uncertain nonlinear systems, *IEEE Transactions on Automatic Control*, Vol. 49, No. 7, pp. 1206−1211, 2004.
[12] P. M. Patre, W. MacKunis, K. Kaiser, W. E. Dixon, Asymptotic tracking for uncertain dynamic systems via a multilayer neural network feedforward and RISE feedback control structure, *IEEE Transactions on Automatic Control*, Vol. 53, No. 9, pp. 2180−2185, 2008.
[13] K. Dupree, P. M. Patre, Z. D. Wilcox, W. E. Dixon, Asymptotic optimal control of uncertain nonlinear Euler–Lagrange systems, *Automatica*, Vol. 47, No. 1, pp. 99−107, 2011.
[14] F. W. Lewis, S. Jagannathan, A. Yesildirak, *Neural Network Control Of Robot Manipulators And Non−Linear Systems*: Taylor & Francis, 1998.
[15] F. L. Lewis, L. Kai, A. Yesildirek, Neural net robot controller with guaranteed tracking performance, *IEEE Transactions on Neural Networks*, Vol. 6, No. 3, pp. 703−715, 1995.
[16] S. Sastry, M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*, pp. 58-59, Prentice Hall, 1989.