مجله مهندسی مکانیک مدرس، فروردین 1397، دوره 18، شماره 01، ص‌ص 388-396

ماهنامه علمی پژوهشی

مهندسی مکانیک مدرس

mme.modares.ac.ir

کنترل به روش یادگیری تقویتی پاندول معکوس چهار درجه آزادی

سید مرتضی خوشرو[1]، مجتبی افتخاری[2*]، مهدی افتخاری[3]

1- دانشجوی کارشناسی ارشد، مهندسی مکانیک، دانشگاه شهید باهنر، کرمان
2- استادیار، مهندسی مکانیک، دانشگاه شهید باهنر، کرمان
3- دانشیار، مهندسی کامپیوتر، دانشگاه شهید باهنر، کرمان
* کرمان، صندوق پستی 133-76175، mo.eftekhari@uk.ac.ir

اطلاعات مقاله

مقاله پژوهشی کامل
دریافت: 02 آبان 1396
پذیرش: 05 دی 1396
ارائه در سایت: 29 دی 1396

کلید واژگان:
یادگیری تقویتی
کنترلر LQR
پاندول معکوس چهار درجه آزادی

چکیده

در این مقاله کنترلر خطی درجه دوم (LQR) مقاوم با استفاده از روش یادگیری تقویتی برای پاندول معکوس چهار درجه آزادی طراحی شده است. سیستم ارائه شده متشکل از یک پاندول معکوس چهار درجه آزادی و یک جرم متمرکز در انتهای آن می‌باشد. ابتدای پاندول در صفحه $x-y$ توانایی حرکت در جهت های x و y را دارد. برای کنترل دو زاویه پاندول معکوس، دو نیروی صفحه ای در جهت های x و y به پایین پاندول وارد می شود. معادلات مدل حاکم بر سیستم با استفاده از روش لاگرانژ استخراج شده اند و سپس یک کنترلر LQR مقاوم بر اساس روش یادگیری تقویتی برای این مسئله طراحی شده است. پاندول برای بازه ای از زاویه های مختلف، طول ها و جرم های مختلف آموزش داده شده است. نامعینی های پارامتری به صورت طول و جرم های مختلف پاندول معکوس و اغتشاشات به صورت نیرو های ضربه ای و متغیر با زمان اعمال شده به پاندول تعریف شده است. پس از یادگیری کنترلر، کنترلر یادگیر می تواند به صورت آنلاین برای بازه ای متفاوت از طول و جرم که قبلا آموزش نیافته و در برابر اغتشاشات پیوسته و ضربه ای که به سیستم اعمال می شود سیستم را کنترل کند. نتایج عددی نشان دهنده عملکرد خوب کنترلر یادگیر در حضور نامعینی های ساختاری و پارامتری، اغتشاشات ضربه ای و پیوسته و نویز سنسورها می‌باشد.

Reinforcement learning control of four degree of freedom inverted pendulum

Seyed Morteza Khoshroo[1], Mojtaba Eftekhari[1*], Mahdi Eftekhari[2]

1- Department of Mechanical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran
2- Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran
* P.O.B. 133-76175 Kerman, Iran, mo.eftekhari@uk.ac.ir

ARTICLE INFORMATION

Original Research Paper
Received 24 October 2017
Accepted 26 December 2017
Available Online 19 January 2018

Keywords:
Reinforcement learning
LQR controller
Four degree of freedom inverted pendulum

ABSTRACT

In this paper, a robust linear quadratic regulator (LQR) based Reinforcement learning method is designed for a four degree of freedom inverted pendulum. The considered system contains a four degree of freedom inverted pendulum with a concentrated mass at the tip of it. The bottom of inverted pendulum is moved in $x-y$ plane in x and y directions. For tracking control of two angles of inverted pendulum, two plane forces are applied in x and y directions at the bottom of pendulum. The governing equations of the system are derived using the Lagrange method and then a robust linear quadratic regulator (LQR) based Reinforcement learning controller is designed. The inverted pendulum is learned for a range of different angles, different lengths and different masses. The parametric uncertainties are defined as various lengths and masses of inverted pendulum and the disturbances are defined as impact and continuous forces which are applied on the inverted pendulum. After learning, the controller can learn online the system for any arbitrary angle, length, mass or disturbance which are not learned in the defined range. Numerical results show that the good performance of the reinforcement learning controller for the inverted pendulum in the presence of structural and parametric uncertainties, impact and continuous disturbances and sensor noises.

1- مقدمه

در مباحث امروزی، کنترل مقاوم جزو مسائلی می‌باشد که در سیستم ها بسیار مفید واقع می شود. مقاوم بودن کنترلر یعنی اینکه سیستم بتواند نسبت به اغتشاشات و نامعینی هایی که امکان دارد در سیستم رخ دهد بهترین عکس العمل را از خود نشان دهد تا سیستم کنترل شود. یکی از روش هایی که می‌توان برای کنترل سیستم ها از آن استفاده کرد، روش آموزش دادن به کنترلر است به گونه‌ای که کنترلر یاد بگیرد که در شرایط مختلف چگونه عمل کند. یکی از این روش‌های آموزش سیستم، روش یادگیری تقویتی[1] می‌باشد. در مسائل یادگیری تقویتی با عاملی روبرو هستیم که از طریق سعی و خطا، با محیط تعامل کرده و یاد می‌گیرد تا عملی بهینه را برای رسیدن به هدف انتخاب کند. با استفاده از این ویژگی، سیستم ها می‌توانند در شرایط مختلف آزمایش شوند تا به مرور توانایی خود را برای تصمیم‌گیری در شرایط محیطی و کاری مختلف افزایش دهند.

[1] Reinforcement learninga

Please cite this article using:
S. M. Khoshroo, M. Eftekhari, M. Eftekhari, Reinforcement learning control of four degree of freedom inverted pendulum, *Modares Mechanical Engineering*, Vol. 18, No. 01, pp. 388-396, 2018 (in Persian)

برای ارجاع به این مقاله از عبارت ذیل استفاده نمایید:

از نمونه کارهایی که در ارتباط با روش یادگیری تقویتی انجام شده است، می‌توان به مقالات [1-3] اشاره کرد.

سیستمی که در این مقاله مورد بررسی قرار گرفته است، یک پاندول معکوس چهار درجه آزادی می‌باشد که در انتهای آن جرم متمرکزی قرار دارد. پاندول معکوس جزو مسائلی است که روش‌های مختلفی برای حل و کنترل آن ارائه می شود. در ادامه به ارائه تحقیقاتی که تا کنون در این زمینه انجام شده است پرداخته‌ایم.

هاواری و همکاران در سال 2006 از کنترل فازی تطبیقی[1] برای کنترل پاندول معکوس بر روی ارابه در حضور اغتشاش استفاده کرده اند. در این مقاله، نمونه آزمایشگاهی ساخته شده و نتایج ناشی از آن نشان‌دهنده کارایی کنترلر فازی می‌باشد [4]. وانگ در سال 2012 فرض کرده است که پاندول معکوس در راستای محورهای x و z حرکت می‌کند و با روش کنترل مود لغزشی[2] سیستم را کنترل کرده، سپس نتایج بدست آمده را با نتایج کنترلر تناسبی-انتگرالی-مشتق گیر[3] مقایسه کرده است [5]. برسیلا و سانکارانارایان در سال 2015 با استفاده از کنترلر غیرخطی[4] پاندولی را کنترل کرده اند که برروی یک چرخ متحرک قرار دارد. از اصطکاک چرخ ها با زمین صرفنظر شده و پاندول با استفاده از گشتاورهای اعمالی به چرخ ها به عنوان ورودی، کنترل شده است [6]. کاسانوا و همکاران در سال 2016 پاندول معکوس را با استفاده از یک نیرو محرکه دورانی کنترل کرده و نتایج را با ساخت مدل بدست آورده اند [7]. خارولا و همکاران در سال 2016 کنترل پاندول معکوس متصل به گاری برروی سطح شیبدار را با استفاده از روش های کنترل فازی[5] و کنترلر تناسبی-انتگرالی-مشتق گیر انجام داده اند. نتایج بدست آمده حاکی از عملکرد بهتر کنترلر تناسبی-انتگرالی-مشتق گیر می‌باشد [8]. روس و همکاران در سال 2017 کنترل پاندول معکوس بروی یگ گاری را با کنترلر جبران ساز توزیع شده موازی فازی[6] و کنترلر تناسبی-انتگرالی-مشتق گیر انجام داده اند. کنترلر فازی بر اساس مدل تاکاگی-سوگینو[7] بوده و نتایج ناشی از آن نشان دهنده عملکرد خوب این کنترلر می‌باشد [9]. گومام و مونیف در سال 2017 با استفاده از کنترل پیشبینی[8]، پاندول معکوس در حضور تاخیر زمانی را کنترل کردند. قانون کنترلی با ترکیب تکنیک اشباع توزیع شده[9] و برگشت به عقب[10] طراحی شده و عملکرد موثر این کنترلر در نتایج نشان داده شده است [10].

با توجه به تحقیقات نام برده شده، می توان گفت تفاوتی که پاندول ارائه شده در این مقاله نسبت به مقالات دیگر دارد، تعداد درجات آزادی آن می‌باشد. در این مسئله، پاندول دارای دو درجه آزادی برای حرکت پایه پاندول در صفحه $x-y$ و دو درجه آزادی برای تغییر زاویه حول محورهای x و y می‌باشد. در نتیجه پاندول در فضای سه بعد توانایی رها شدن در هر سمتی را دارد. همچنین کنترلر ارائه شده با استفاده از روش یادگیری تقویتی طراحی شده است که در کارهای قبلی از این روش استفاده نشده است.

ابتدا برای اعتبار سنجی روش یادگیری تقویتی، برای سیستم پاندول معکوس و جرم متمرکز دو درجه آزادی، نتایج برتری کنترلر یادگیری تقویتی نسبت به روش LQR ارائه شده است، سپس کنترلر با استفاده از روش یادگیری تقویتی برای سیستم پاندول معکوس و جرم متمرکز چهار درجه آزادی طراحی شده و نتایج با کنترلر LQR مقایسه و ارائه شده اند. ابتدا کنترلر برای حداقل محدوده مورد نیاز برای سیستم آموزش، سپس در محدوده های مختلف که کنترلر در آنها آموزشی ندیده است اجرا شده است تا محدوده توانایی کنترلر برای کنترل سیستم مشخص شود. برای مقایسه دو کنترلر برای پاندول چهار درجه آزادی علاوه بر آزمایش کنترلرها برای شرایط انحراف زاویه اولیه و جایجایی اولیه بزرگ پایه پاندول، مقاومت کنترلر در برابر نامعینی های پارامتری و ساختاری، اغتاشات ضربه ای وغیر ضربه ای و همچنین نویز سنسورها بررسی و پاسخ ها رسم شده اند. نتایج بدست آمده نشان دهنده این می‌باشد که کنترلر یادگیری تقویتی نسبت به کنترلر LQR در شرایط در نظر گرفته شده بهتر عمل می کند.

2- طرح مسأله و معادلات حاکم

سیستمی که در این مقاله مورد بررسی قرار گرفته شده است، تشکیل می شود از یک پاندول معکوس با طول L و جرم Ml که در انتهای آن جرم متمرکزی با جرم M قرار دارد. پاندول معکوس دارای چهار درجه آزادی می‌باشد که دو درجه آزادی برای تغییر زاویه و دو درجه آزادی مربوط به تغییر موقعیت مکانی ابتدای پاندول است. زاویه های انحراف پاندول از حالت تعادل برای هر یک از درجات آزادی، θ و φ می باشند. و موقعیت مکانی ابتدای پاندول معکوس در صفحه با دو پارامتر x و y مشخص می شود. جابجایی ابتدای پاندول توسط دو نیروی F_x و F_y وارد شده به آن صورت می گیرد. شکل 1.

معادلات مدل ریاضی حاکم بر سیستم پاندول معکوس و جرم متمرکز با استفاده از روش لاگرانژ استخراج شده است و کنترلر طراحی شده برای این سیستم، کنترلر خطی درجه دوم (LQR) مقاوم با استفاده از روش یادگیری تقویتی می‌باشد. در ادامه به توضیح کامل معادلات دینامیکی و کنترلی سیستم پرداخته شده است.

2-1- دینامیک لاگرانژ

برای سیستم غیر مقید[11]، معادله لاگرانژ بصورت زیر تعریف می شود:

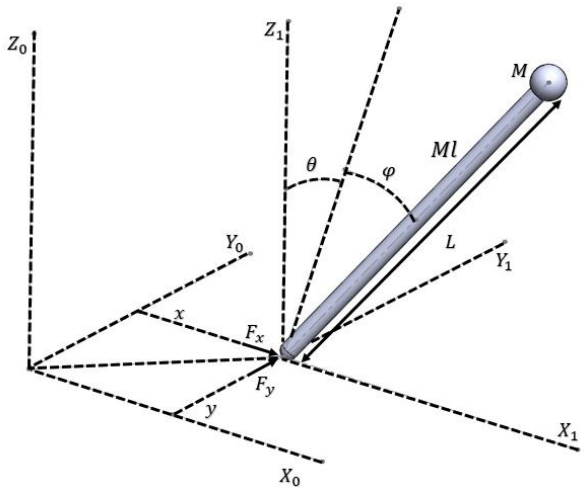


Fig.1 Schematic of four degree of freedom of inverted pendulum with concentrated mass

شکل 1 شکل شماتیکی پاندول معکوس چهار درجه آزادی با جرم متمرکز

[1] Adaptive fuzzy controller
[2] Sliding mode controller
[3] PID controller
[4] Nonlinear controller
[5] Fuzzy controller
[6] Fuzzy parallel distributed compensation controller
[7] Takagi-Sugeno
[8] Predictor-based control
[9] Nested saturation
[10] Back stepping
[11] Nonholonomic system

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right)-\frac{\partial L}{\partial q_i}=Q_i \tag{1}$$

که q_i و Q_i به ترتیب مختصات و نیروهای تعمیم یافته برای سیستم می باشند. همچنین لاگرانژین (L)، تفاضل انرژی جنبشی و پتانسیل کل سیستم است.

برای سیستم پاندول معکوس و جرم متمرکز، مختصات و نیروهای تعمیم یافته عبارت اند از:

$$q=[x,y,\theta,\varphi]^{\mathrm{T}} \tag{2}$$

$$Q=[F_x\,,F_y,0\,,0]^{\mathrm{T}} \tag{3}$$

حال برای محاسبه انرژی جنبشی و پتانسیل کل سیستم با توجه به شکل 2، بردارهای موقعیت مرکز جرم ها و ابتدای پاندول معکوس به صورت زیر در می آیند:

$$\vec{b}^{0}=\begin{bmatrix}x\\y\\0\end{bmatrix} \tag{4}$$

$$\vec{m}_1^{1}=\begin{bmatrix}0\\0\\\frac{L}{2}\end{bmatrix} \tag{5}$$

$$\vec{m}_2^{1}=\begin{bmatrix}0\\0\\L\end{bmatrix} \tag{6}$$

$$\vec{m}_1^{0}=\vec{b}^{0}+\vec{R}_1^{0}\vec{m}_1^{1} \tag{7}$$

$$\vec{m}_2^{0}=\vec{b}^{0}+\vec{R}_1^{0}\vec{m}_2^{1} \tag{8}$$

که در معادلات 7 و 8، $\vec{R}_1^{0}$ ماتریس دوران دستگاه متصل به پاندول معکوس و جرم متمرکز (x_1,y_1,z_1) نسبت به دستگاه پایه (x_0,y_0,z_0) می‌باشد که بصورت زیر تعریف می شود:

$$\vec{R}_1^{0}=\begin{bmatrix}1&0&0\\0&\cos\theta&-\sin\theta\\0&\sin\theta&\cos\theta\end{bmatrix}\begin{bmatrix}\cos\varphi&0&\sin\varphi\\0&1&0\\-\sin\varphi&0&\cos\varphi\end{bmatrix} \tag{9}$$

در ادامه با جایگزینی با استفاده از معادلات مرکز جرم، انرژی پتانسیل سیستم حاصل می شود:

$$U_1=\vec{G}^{0\,\mathrm{T}}Ml\,\vec{m}_1^{0} \tag{10}$$

$$U_2=\vec{G}^{0\,\mathrm{T}}M\,\vec{m}_2^{0} \tag{11}$$

حال سرعت خطی مرکز جرم ها با مشتقگیری زمانی از معادلات 7 و 8:

$$\vec{V}_1^{0}=\frac{d}{dt}(\vec{m}_1^{0}) \tag{12}$$

$$\vec{V}_2^{0}=\frac{d}{dt}(\vec{m}_2^{0}) \tag{13}$$

و سرعت زاویه ای مرکز جرم ها نیز با توجه به ماتریس دوران دستگاه یک نسبت به صفر (معادله 9)، بدست می آید:

$$\vec{w}_1^{0}=\frac{d}{dt}(\vec{R}_1^{0})\vec{R}_1^{0\,\mathrm{T}} \tag{14}$$

ممان اینرسی پاندول معکوس و جرم متمرکز با استفاده از معادله 9، نسبت به دستگاه پایه بصورت زیر نوشته می شوند:

$$\vec{I}_1^{0}=\vec{R}_1^{0}\vec{I}_1^{1}\vec{R}_1^{0\,\mathrm{T}} \tag{15}$$

$$\vec{I}_2^{0}=\vec{R}_1^{0}\vec{I}_2^{1}\vec{R}_1^{0\,\mathrm{T}} \tag{16}$$

در نتیجه انرژی جنبشی هریک از مراکز جرم بدست می آید:

$$T_1=\frac{1}{2}\vec{V}_1^{0\,\mathrm{T}}Ml\,\vec{V}_1^{0}+\frac{1}{2}\,\vec{w}_1^{0\,\mathrm{T}}\vec{I}_1^{0}\,\vec{w}_1^{0} \tag{17}$$

$$T_2=\frac{1}{2}\,\vec{V}_2^{0\,\mathrm{T}}M\,\vec{V}_2^{0}+\frac{1}{2}\,\vec{w}_2^{0\,\mathrm{T}}\vec{I}_2^{0}\,\vec{w}_2^{0} \tag{18}$$

حال با توجه به انرژی پتانسیل و جنبشی بدست آمده برای کل سیستم، لاگرانژین برای سیستم پاندول معکوس و جرم متمرکز برابر است با:

$$\begin{aligned}L=&\frac{M\,\dot{x}^2}{2}+\frac{M\,\dot{y}^2}{2}+\frac{Ml\,\dot{x}^2}{2}+\frac{Ml\,\dot{y}^2}{2}\\&+L^2\,M\dot{\varphi}^2+\frac{7\,L^2\,Ml\,\dot{\varphi}^2}{24}+\,L^2\,M\,\dot{\theta}^2\cos^2(\varphi)\\&+\frac{7\,L^2\,Ml\,\dot{\theta}^2\cos^2(\varphi)}{24}+L\,M\,\dot{\varphi}\,\dot{x}\cos(\varphi)\\&+\frac{L\,Ml\,\dot{\varphi}\,\dot{x}\cos(\varphi)}{2}-L\,M\,g\cos(\varphi)\cos(\theta)\\&-\frac{L\,Ml\,g\cos(\varphi)\cos(\theta)}{2}\\&-L\,M\,\dot{\theta}\,\dot{y}\cos(\varphi)\cos(\theta)\\&-\frac{L\,Ml\,\dot{\theta}\,\dot{y}\cos(\varphi)\cos(\theta)}{2}\\&+L\,M\,\dot{\varphi}\,\dot{y}\sin(\varphi)\sin(\theta)\\&+\frac{L\,Ml\,\dot{\varphi}\,\dot{y}\sin(\varphi)\sin(\theta)}{2}\end{aligned} \tag{19}$$

در آخر با فاکتورگیری نسبت به مشتق دوم مختصات تعمیم یافته، معادله دینامیکی سیستم به فرم زیر در می آید:

$$M_T\ddot{q}+N_T=F_T \tag{20}$$

که در این معادله پارامترهای $\ddot{q}=[\ddot{x},\ddot{y},\ddot{\theta},\ddot{\varphi}]^{\mathrm{T}}$، $F_T=[F_x\,,F_y]^{\mathrm{T}}$ و $M_T\in\mathbb{R}^{4\times4}$ و $N_T\in\mathbb{R}^{4\times1}$ می باشند.

2-2- الگوریتم یادگیری تقویتی

کنترلر یادگیری تقویتی، سیستم را در شرایط مختلف آزمایش و سعی می کند تا با محیط تعامل کرده و یاد بگیرد که عملی بهینه را برای رسیدن به هدف انتخاب کند. برای این منظور زمانی که سیستم در یک شرایط خاص آزمایش می گیرد، کنترلر سیستم را در همان شرایط در دفعات مختلف آزمایش و در هر تکرار سعی میکند تا نیروهای وارده به سیستم را بهبود بخشد تا سیستم کنترل شود.

در ادامه برای طراحی کنترلر LQR با روش یادگیری تقویتی، از تعاریف مختلفی استفاده شده است که با استفاده از مرجع [11]، به ترتیب توضیح داده می شوند. ابتدا دینامیک سیستم در فرایند تصمیم گیری مارکو افق محدود[13]، تعریف می شود که این فرایند شامل پارامترهای زیر می‌باشد:

S : مجموعه متغیرهای حالت[14]

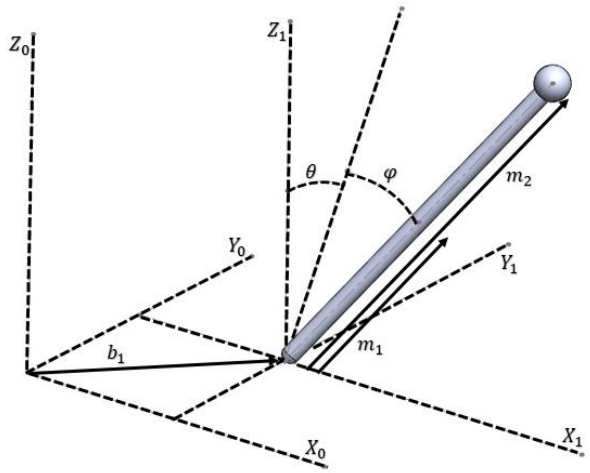


Fig.2 Vector position of center of masses and coordinate of inverted pendulum with concentrated mass

شکل 2 موقعیت بردار مرکز جرم ها و مختصات پاندول معکوس با جرم متمرکز

[13] Finite horizon Markov decision process
[14] Set of states variables

a : مجموعه نیروها[1]

P_{sa} : احتمال وقوع حالت[2]

R : تابع پاداش[3]

T : افق زمانی[4]

برای سیستم پاندول معکوس، مجموعه متغیرهای حالت (s) برابرند با پارامترهای حالت سیستم، و مجموعه نیروها (a) برابرند با نیروهای کنترلر که در سیستم کار انجام می دهند. در نتیجه داریم:

$$s = [x, y, \theta, \varphi, \dot{x}, \dot{y}, \dot{\theta}, \dot{\varphi}]^{\mathrm{T}} \quad (21)$$

$$a = [F_x, F_y]^{\mathrm{T}} \quad (22)$$

در سیستم های مکانیکی که دینامیک مسئله بیانگر حالات سیستم می‌باشد، احتمال وقوع حالت (P_{sa}) بر اساس حل عددی دینامیک سیستم تعریف می شود:

$$S_{t+1} = S_t + \Delta t\, \dot{S}_t \triangleq f(s_t, a_t) \quad (23)$$

در معادله 23، اندیس t نشانگر زمان و Δt تغییرات زمانی می‌باشد، در نتیجه حالت بعدی سیستم در زمان $t+1$ از معادله بالا بدست می آید.

تابع پاداش (R) با استفاده از رگرسیون خطی درجه دوم[5] تعریف می‌شود:

$$R(s_t, a_t) = -(\, s_t{}^{\mathrm{T}}\, U\, s_t + a_t{}^{\mathrm{T}}\, V\, a_t\,) \quad (24)$$

در معادله بالا $U \in \mathbb{R}^{8\times 8}$ و $V \in \mathbb{R}^{2\times 2}$ می باشند.

علاوه بر استفاده از معادله خطی در تابع پاداش، معادله دینامیکی سیستم نیز خطی سازی می شود. در نتیجه با استفاده از معادله خطی سازی بسط تیلور:

$$S_{t+1} \approx f(\bar{s}_t, \bar{a}_t) + \left(\nabla_s f(\bar{s}_t, \bar{a}_t)\right)^{\mathrm{T}}(s_t - \bar{s}_t) + \left(\nabla_a f(\bar{s}_t, \bar{a}_t)\right)^{\mathrm{T}}(a_t - \bar{a}_t) \quad (25)$$

سیستم را حول نقاط تعادل $\bar{s}$ و $\bar{a}$ که نقاط تعادل سیستم می باشند خطی سازی می کنیم. در نتیجه معادله 23 بصورت زیر در می آید:

$$s_{t+1} \approx A_t\, s_t + B_t\, a_t + D_t\, w_t \quad (26)$$

که $A \in \mathbb{R}^{8\times 8}$، $B \in \mathbb{R}^{8\times 2}$ و $D \in \mathbb{R}^{8\times 4}$ می باشند.

در معادله 26، w اغتشاشات وارد شده به سیستم، بصورت نیروی ضربه ای وارد شده به ابتدای پاندول معکوس و جرم متمرکز تعریف می شوند:

$$w = [\overline{F_x}, \overline{F_y}, \overline{F_\theta}, \overline{F_\varphi}]^{\mathrm{T}} \quad (27)$$

در معادله 27، $\overline{F_x}$ و $\overline{F_y}$ نیروهای ضربه ای وارد شده به ابتدای پاندول که باعث تغییر موقعیت ابتدای پاندول معکوس و $\overline{F_\theta}$ و $\overline{F_\varphi}$ نیروهای ضربه ای وارد شده به جرم متمرکز می باشند که باعث تغییر زاویه پاندول می شوند.

همانطور که قبلا گفته شد در روش یادگیری تقویتی نیروها باید با تکرار و آزمایش بدست آیند. در نتیجه باید مجموعه نیروها (a) بصورت تابعی از دیگر پارامترهای تعریف شده در روش مارکو نوشته شوند. برای این منظور با استفاده از روش تکرار ارزش[6] و تعریف پارامترهای تابع ارزش[7] (V^*) و سیاست[8] (π^*) به صورت زیر:

$$V^*(s_t) = \max_{a_t} R(s_t, a_t) + \mathrm{E}_{s_{t+1}\sim P_{s_t a_t}}[V^*(s_{t+1})] \quad (28)$$

$$\pi^*(s_t) = \operatorname{argmax}_{a_t} R(s_t, a_t) + \mathrm{E}_{s_{t+1}\sim P_{s_t a_t}}[V^*(s_{t+1})] \quad (29)$$

معادله نیرو که به معادله ریکاتی[9] معروف است، با توجه به روند اثبات پایداری لیاپانوف که در قسمت بعد آورده شده است، به صورت زیر بدست می‌آید:

$$a_t = -\left(B_t^{\mathrm{T}}\varphi_{t+1}B_t - V_t\right)^{-1}B_t^{\mathrm{T}}\varphi_{t+1}A_t s_t \quad (30)$$

$$\varphi_t = A_t^{\mathrm{T}} \left(\varphi_{t+1} - \varphi_{t+1}B_t\left(B_t^{\mathrm{T}}\varphi_{t+1}B_t - V_t\right)^{-1}B_t^{\mathrm{T}}\varphi_{t+1}\right)A_t - U_t \quad (31)$$

حال با توجه به پارامترهای تعریف شده سیستم باید یاد بگیرد که در هر شرایطی چگونه عمل کند. روند یادگیری سیستم با الگوریتم برنامه ریزی دینامیک تفاضلی[10]، انجام می شود که مراحل الگوریتم به ترتیب بصورت زیر می باشند:

1- ابتدا برای سیستم یک شرایط اولیه s_0 و a_0 انتخاب می شود و با استفاده از معادله 23 و انتخاب a_t، حالت های بعدی سیستم را تا زمان محدود T بدست می آوریم.

2- سیستم حول نقاط تعادل خطی سازی می شود تا معادله 26 حاصل شود.

3- با استفاده از معادلات 30، 31 و حالت های بدست آمده از قسمت شماره 1 نیروهای جدید محاسبه می شوند.

4- حال با استفاده از نیروهای جدید بدست آمده از قسمت شماره 3 و شرایط اولیه s_0، همانند قسمت اول، حالت های بعدی سیستم را محاسبه می کنیم و روند تکرار قسمت های 1 تا 4 را تا زمانی که تمام نیروها به بهترین میزان خود برسند ادامه می دهیم.

در ادامه، الگوریتم گفته شده را در شرایط اولیه مختلف برای پاندول معکوس و جرم متمرکز اجرا می کنیم. کنترلر برای هر شرط اولیه بهترین عکس العمل را برای تعادل پاندول بدست می آورد و در حافظه خود نگاه می دارد. در نتیجه پس از اتمام فرایند یادگیری برای شرایط مختلف تعریف شده برای سیستم، کنترلر بصورت زیر عمل می کند: مقدار زاویه ای که سیستم در آن قرار دارد را با مقادیر زاویه ای که یاد گرفته است مقایسه و نزدیکترین حالت به حالت فعلی سیستم را انتخاب می کند. سپس با استفاده از معادله ای شبیه به معادله 30، تصحیحی برروی نیرو انجام می دهد و نیرو را اعمال می کند.

2-3- اثبات پایداری کنترلر یادگیری تقویتی

ابتدا تابع لیاپانوف را به صورت منفی تابع ارزش در نظر می گیریم که یک تابع مثبت معین می‌باشد:

$$V = -V^*(s_t) \quad (32)$$

زیرا فرم انتخاب شده برای تابع پاداش (معادله 24)، همواره کوچکتر مساوی با صفر است. با مشتق گیری از تابع لیاپانوف می توان نوشت:

$$\dot{V} = -\dot{V}^*(s_t) = -\frac{V^*(s_{t+1}) - V^*(s_t)}{\Delta t} = -\frac{s_{t+1}^{\mathrm{T}}\varphi_{t+1}s_{t+1} - s_t^{\mathrm{T}}\varphi_t s_t}{\Delta t} \quad (33)$$

با جایگذاری $s_{t+1} = A_t\, s_t + B_t\, a_t$ در معادله بالا و همچنین جایگذاری a_t و φ_t از معادلات 30 و 31 در معادله 33، به رابطه زیر می‌رسیم:

$$\dot{V} = R(s_t, a_t) \le 0 \quad (34)$$

از آنجا که تابع $R(s_t, a_t)$ همواره منفی است در نتیجه سیستم پایدار است.

[15] Set of actions
[16] State transition distribution
[17] Reward function
[18] Horizon time
[19] Linear quadratic regulation
[20] Value iteration
[21] Value function
[22] Policy
[23] Riccati equation
[24] Differential dynamic programing

3- نتایج و بحث

تمامی محاسبات مربوط به مدل ریاضی دینامیک پاندول معکوس و جرم متمرکز و همچنین کنترلر طراحی شده برای این سیستم، در نرم افزار متلب کدنویسی و نتایج حاصل از کنترل سیستم نیز در این نرم افزار رسم شده اند.

ابتدا برای اعتبار سنجی کنترلر LQR با استفاده از روش یادگیری تقویتی، کنترلر برای پاندول دو درجه آزادی که در مقالات متنوع بررسی و حل شده است، طراحی و همچنین نتایج به دست آمده با کنترلر LQR مقایسه شده است. هر دو کنترلر برای سیستم پاندول معکوس و جرم متمرکز دو درجه آزادی، در شرایط دینامیکی و شماتیکی یکسان طراحی و در شرایط مختلف آزمایش شده اند. نتایج حاصل از این مقایسه برای دو حالت بررسی شده و به ترتیب در شکل های 3-الف، 3-ب و 3-ج آورده شده است که عبارت اند از: 1- پاسخ زمانی سیستم به ازای زاویه انحراف اولیه و جابجایی پایه اولیه متوسط در شکل 3-الف، 2- پاسخ زمانی سیستم به ازای زاویه انحراف اولیه بزرگ و جابجایی پایه اولیه صفر در شکل 3-ب و 3- پاسخ زمانی سیستم به ازای زاویه انحراف اولیه صفر و جابجایی پایه اولیه بزرگ در شکل 3-ج. برای هر یک از پاسخ های زمانی سیستم، فرض شده است که مقدار پارامترهای اولیه انتخاب شده برای این سیستم عبارتند از: $L = 35\ \text{cm}, Ml = 150\ \text{g}, M = 100\ \text{g}$. در شکل 3-الف پاسخ زمانی هر دو کنترلر برای $x = 2\ \text{m}, \varphi = 20^\circ$ رسم شده است. همانطور که مشخص می‌باشد کنترلر LQR با استفاده از یادگیری تقویتی با سرعت بیشتر، نیرو و تغییرات نوسانی کمتر نسبت به LQR سیستم را به حالت تعادل رسانیده است. در شکل 3-ب پاسخ زمانی هر دو کنترلر برای $x = 0\ \text{m}, \varphi = 40^\circ$ رسم شده است. همانطور که مشخص می‌باشد کنترلر LQR با استفاده از یادگیری تقویتی در کمتر از سه ثانیه سیستم را به حالت تعادل رسانیده است در حالیکه LQR نتوانسته است سیستم را به تعادل برساند. در شکل 3-ج پاسخ زمانی هر دو کنترلر برای $x = 4\ \text{m}, \varphi = 0^\circ$ رسم شده است. همانطور که مشخص می‌باشد کنترلر LQR با استفاده از یادگیری تقویتی در کمتر از چهار ثانیه سیستم را به حالت تعادل رسانیده است در حالیکه LQR نتوانسته است سیستم را به تعادل برساند.

حال با توجه به عملکرد بهتر کنترلر LQR با استفاده از یادگیری تقویتی، این کنترلر برای سیستم پاندول معکوس و جرم متمرکز چهار درجه آزادی طراحی و سپس آموزش دیده است. در ادامه به نتایج بدست آمده حاصل از پاسخ های زمانی این کنترلر و مقایسه آن با کنترلر LQR پرداخته شده است.

ابتدا کنترلر در حداقل بازه ممکن آموزش دیده است. پارامترها و اندازه هایی که برای آموزش سیستم انتخاب شده اند در جدول شماره 1 آورده شده است. با توجه به جدول 1، می توان اشاره کرد که سیستم در یک جرم و طول یکسان آموزش دیده. کنترلر طراحی شده، یادگیری خود را برای مدت زمان 5 ثانیه با مقادیر تعریف شده در جدول 1، مقادیر $(-30^\circ, -20^\circ, -10^\circ, 0^\circ, 10^\circ, 20^\circ, 30^\circ)$ برای θ و φ و مقدار صفر برای x و y انجام داده است. در نتیجه کنترلر، $7^2 = 49$ مرتبه اجرا شده است تا تمام حالات انتخاب شده برای زاویه ها را پوشش دهد. همچنین دفعات تکرار در هر اجرا برای بدست آوردن بهترین عکس العمل 51 مرتبه می‌باشد. پس از اتمام یادگیری، کنترلر در شرایط مختلف آزمایش و نتایج به دست آمده با کنترلر LQR مقایسه شده و نتایج رسم شده اند.

شکل 4 پاسخ کنترلر LQR با استفاده از یادگیری تقویتی در مقایسه با LQR در سیستم پاندول چهار درجه آزادی به ازای نامعینی پارامتری می‌باشد. شکل 4-الف پاسخ کنترلر LQR با استفاده از یادگیری تقویتی و LQR را برای حالتی که طول پاندول به اندازه 10 درصد طول اولیه، جرم پاندول به اندازه 20 درصد جرم پاندول اولیه و جرم متمرکز به اندازه 30 درصد مقدار جرم متمرکز اولیه افزایش یافته است را نشان می دهد. همانطور که مشخص است کنترلر LQR با استفاده از یادگیری تقویتی در برابر تغییرات پارامتری توانسته است با اعمال نیرو در کمتر از پنج ثانیه موقعیت و زوایا را به صفر برگرداند. در حالی که کنترلر LQR در ابتدا از حالت تعادل خارج می شود. شکل 4-ب پاسخ کنترلر LQR با استفاده از یادگیری تقویتی و LQR را برای حالتی که طول پاندول و جرم پاندول تغییر نکنند و جرم متمرکز به اندازه 60 درصد جرم اولیه خود افزایش یافته باشد را نشان می دهد. همانطور که مشخص است کنترلر LQR با استفاده از یادگیری تقویتی در برابر این نوع تغییرات پارامتری نیز توانسته است با اعمال نیرو در کمتر از پنج ثانیه موقعیت و زوایا را به صفر برگرداند، در حالی که کنترلر LQR در ابتدا از حالت تعادل خارج می شود.

شکل 5 پاسخ کنترلر LQR با استفاده از یادگیری تقویتی در مقایسه با LQR در سیستم پاندول چهار درجه آزادی به ازای نامعینی ساختاری می‌باشد. شکل 5 پاسخ کنترلر LQR با استفاده از یادگیری تقویتی و LQR برای نامعینی ساختاری رسم شده است. نامعینی ساختاری به صورت زیر در نظر گرفته شده است:

$$\dot{X} = f(X) + g(X)u + \Delta f(X) \quad (35)$$

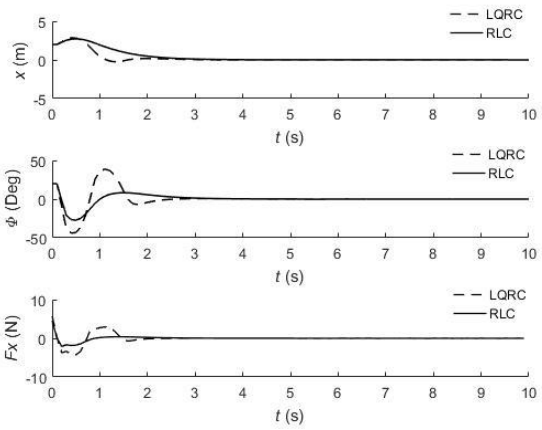
که مقدار $\Delta = 0.3$ انتخاب شده است. همانطور که در شکل 5 نشان داده شده، کنترلر LQR با استفاده از یادگیری تقویتی در کمتر از چهار ثانیه پاندول را به حالت تعادل رسانده است در حالی که کنترلر LQR در همان ابتدا از تعادل خارج می شود.

شکل 6 و 7 پاسخ کنترلر LQR با استفاده از یادگیری تقویتی در مقایسه با LQR در سیستم پاندول چهار درجه آزادی به ازای اغتشاش ضربه ای و غیر ضربه ای می باشند. شکل 6-الف پاسخ سیستم را به ازای نیروی ضربه ای 6 نیوتن در راستای x و y که در ثانیه پنج به سیستم وارد شده است را نشان می دهد. در کنترلر LQR با استفاده از یادگیری تقویتی سیستم به حالت تعادل برگشته در حالی که LQR از ثانیه هفت به بعد از تعادل خارج شده است. شکل 6-ب پاسخ سیستم را به ازای نیروی ضربه ای 0.95 نیوتن در راستاهای θ و φ که در ثانیه پنج به سیستم وارد شده است را نشان می دهد. کنترلر LQR با استفاده از یادگیری تقویتی سیستم را به حالت تعادل برگردانده در حالی که LQR از ثانیه هفت به بعد از تعادل خارج شده است. شکل 7-الف پاسخ سیستم را به ازای نیروی غیر ضربه ای سینوسی (تابع $3\sin t$ نیوتون) که در راستای x و y که بین ثانیه سه تا شش به سیستم وارد شده است را نشان می دهد. در کنترلر LQR با استفاده از یادگیری تقویتی سیستم به حالت تعادل برگشته در حالی که LQR از همان ابتدا از تعادل خارج شده است. شکل 7-ب پاسخ سیستم را به ازای نیروی غیر ضربه ای سینوسی (تابع $0.45\sin t$نیوتون) که در راستای θ و φ که بین ثانیه سه تا شش به سیستم وارد شده است را نشان می دهد. کنترلر LQR با استفاده از یادگیری تقویتی سیستم را به حالت تعادل برگردانده در حالی که LQR از همان ابتدا از تعادل خارج شده است. شکل 8 پاسخ کنترلر LQR با استفاده از یادگیری تقویتی در مقایسه با LQR در سیستم پاندول

جدول 1 پارامترها و مقادیر

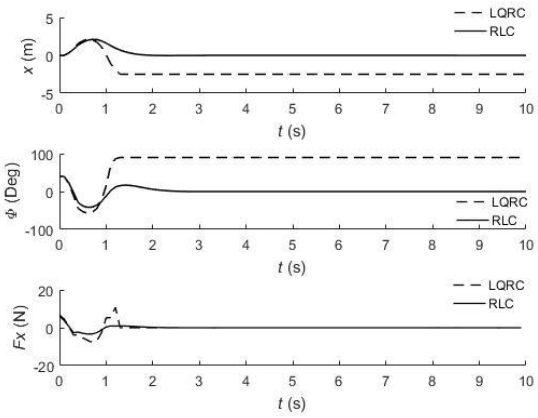
Table 1 Parameters and values

پارامتر	مقدار
L	35 cm
Ml	150 g
M	100 g
Δt	0.1
g	9.8 m/s^2
$\bar{s}$	$[0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0]^{\mathrm{T}}$
$\bar{a}$	$[0\ \ 0]^{\mathrm{T}}$ N
U	$\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
V	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$



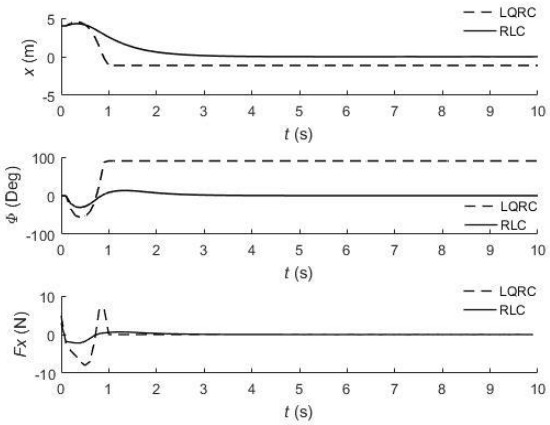
a) System response for the average amount of initial angle and initial displacement of the bottom of the inverted pendulum
$L = 35$ cm , $Ml = 150$ g , $M = 100$ g
$x = 2$ m , $\varphi = 20^{\circ}$

الف) پاسخ سیستم به ازای مقدار انحراف زاویه اولیه و جابجایی اولیه متوسط پایه پاندول



b) System response for the large amount of the initial angle and the zero amount of initial displacement of the bottom of inverted pendulum
$L = 35$ cm , $Ml = 150$ g , $M = 100$ g
$x = 0$ m , $\varphi = 40^{\circ}$

ب) پاسخ سیستم به ازای مقدار انحراف زاویه اولیه بزرگ و جابجایی پایه پاندول اولیه صفر



c) System response for the zero amount of the initial angle and the large amount of initial displacement of the bottom of pendulum
$L = 35$ cm , $Ml = 150$ g , $M = 100$ g
$x = 4$ m , $\varphi = 0^{\circ}$

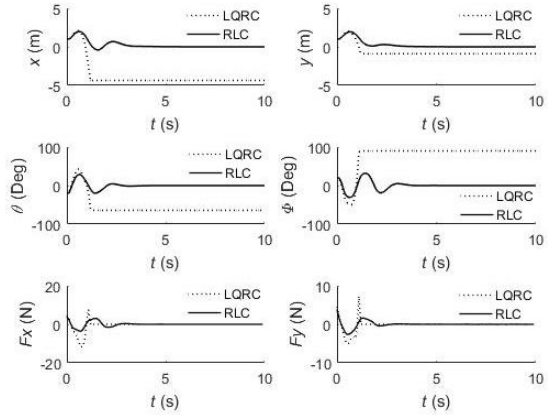
ج) پاسخ سیستم به ازای مقدار انحراف زاویه اولیه صفر و جابجایی پایه پاندول اولیه بزرگ

Fig.3 Time response of Reinforcement learning controller and LQR controller for two-degree of freedom inverted pendulum for different angles and positions

شکل 3 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول دو درجه آزادی برای زوایا و موقعیت های مختلف

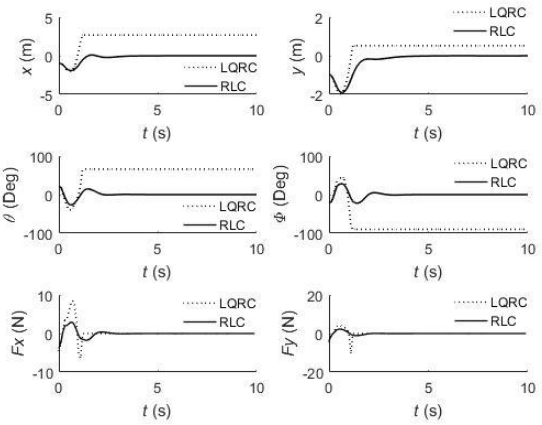
چهار درجه آزادی به ازای اثرات نویز سنسورها می‌باشد. شکل 8-الف پاسخ سیستم را به ازای تابع تصادفی نویز سفید در مقدار دریافتی زاویه توسط سنسورها و شکل 8-ب پاسخ سیستم را به ازای تابع تصادفی نویز سفید در تمام مقادیر دریافتی توسط سنسورها نشان می دهد. در هر دو شکل 8-الف و 8-ب کنترلر LQR با استفاده از یادگیری تقویتی و LQR سیستم را به حالت تعادل بر می گرداند.

شکل 9 و 10 پاسخ کنترلر LQR با استفاده از یادگیری تقویتی در مقایسه با LQR در سیستم پاندول چهار درجه آزادی به ازای زاویه های انحراف اولیه بزرگ و جابجایی اولیه بزرگ پایه پاندول در جهت های x و y می‌باشد. شکل 9 پاسخ سیستم به ازای مقدار زاویه های اولیه 33 درجه برای هر یک از زاویه های θ و φ و مقدار اولیه صفر برای پایه پاندول در راستاهای



a) System response to the increasing of pendulum length, pendulum mass and concentrated mass
$L = 38.5$ cm , $Ml = 180$ g , $M = 130$ g
$x = 1$ m , $y = 1$ m , $\theta = -20^{\circ}$, $\varphi = 20^{\circ}$

الف) پاسخ سیستم نسبت به افزایش طول پاندول، جرم پاندول و جرم متمرکز



b) System response to the increasing of concentrated mass
$L = 35\ \text{cm}\ , Ml = 150\ \text{g}\ , M = 160\ \text{g}$
$x = -1\ \text{m}\ , y = -1\ \text{m}\ , \theta = 20^{\circ}, \varphi = -20^{\circ}$

ب) پاسخ سیستم نسبت به افزایش مقدار جرم متمرکز

Fig.4 Time response of Reinforcement learning controller and LQR controller for four-degree of freedom inverted pendulum for parametric uncertainties

شکل 4 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول چهار درجه آزادی برای نامعینی های پارامتری

x و y می‌باشد. همانطور که در شکل 9 مشخص می‌باشد کنترلر LQR با استفاده از یادگیری تقویتی سیستم را به حالت تعادل رسانیده است در حالی که LQR به حالت نود درجه در آمده است. شکل 10 پاسخ سیستم به ازای مقدار زاویه های اولیه صفر درجه برای هر یک از زاویه های θ و φ و مقدار اولیه 3.3 متر برای پایه پاندول در راستاهای x و y می‌باشد. همانطور که در شکل 10 مشخص می‌باشد کنترلر LQR با استفاده از یادگیری تقویتی سیستم را به حالت تعادل رسانیده است در حالیکه LQR به حالت نود درجه در آمده است.

4- نتیجه گیری

در این مقاله از کنترلر LQR مقاوم بر اساس روش یادگیری تقویتی برای

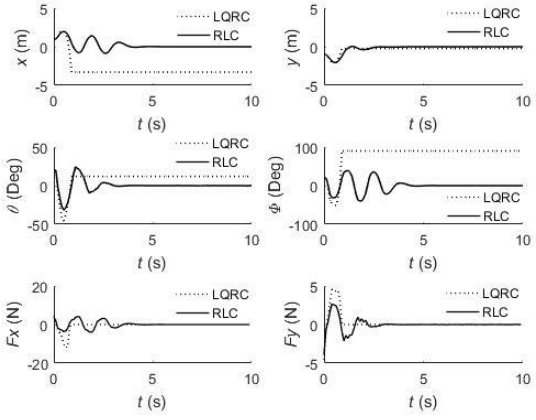
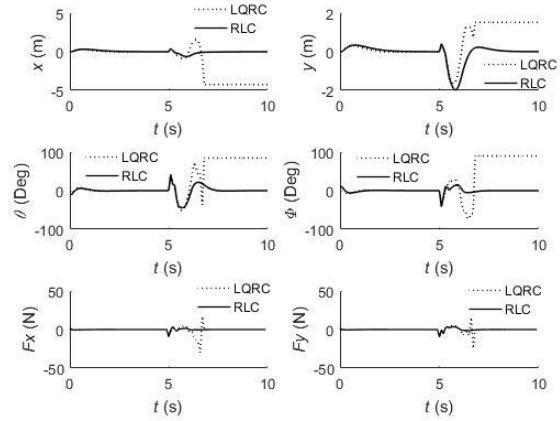


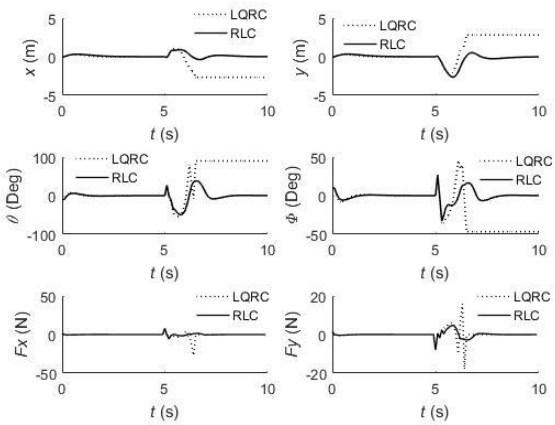
Fig.5 Time response of Reinforcement learning controller and LQR controller for four-degree of freedom inverted pendulum for structural uncertainty
$L = 35\ \text{cm}\ , Ml = 150\ \text{g}\ , M = 100\ \text{g}$
$x = 1\ \text{m}\ , y = -1\ \text{m}\ , \theta = 20^{\circ}, \varphi = 20^{\circ}$

شکل 5 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول چهار درجه آزادی برای نامعینی ساختاری



a) The response of the system to the impact force in the x and y directions
$L = 35\ \text{cm}\ , Ml = 150\ \text{g}\ , M = 100\ \text{g}$
$x = 0\ \text{m}\ , y = 0\ \text{m}\ , \theta = -10^{\circ}, \varphi = 10^{\circ}$
$t = 5\ \text{s}$ در $\overline{F_x} = 6\ \text{N}$
$t = 5\ \text{s}$ در $\overline{F_y} = 6\ \text{N}$

الف) پاسخ سیستم نسبت به نیروی ضربه ای در راستاهای x و y



b) The response of the system to the impact force in the θ and φ directions
$L = 35\ \text{cm}\ , Ml = 150\ \text{g}\ , M = 100\ \text{g}$
$x = 0\ \text{m}\ , y = 0\ \text{m}\ , \theta = -10^{\circ}, \varphi = 10^{\circ}$
$t = 5\ \text{s}$ در $\overline{F_\theta} = 0.95\ \text{N}$
$t = 5\ \text{s}$ در $\overline{F_\varphi} = 0.95\ \text{N}$

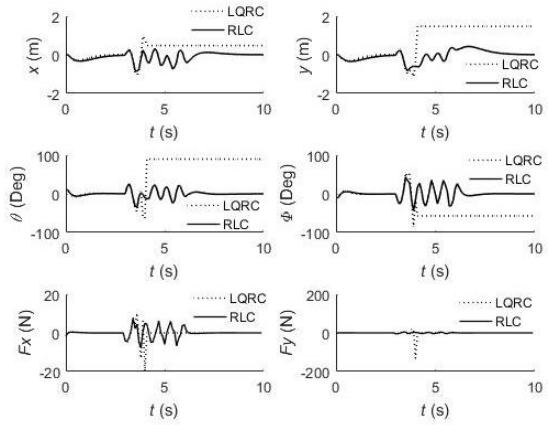
ب) پاسخ سیستم نسبت به نیروی ضربه ای در راستاهای θ و φ

Fig.6 Time response of Reinforcement learning controller and LQR controller for four-degree of freedom inverted pendulum for impact disturbance

شکل 6 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول چهار درجه آزادی برای اغتشاش ضربه ای

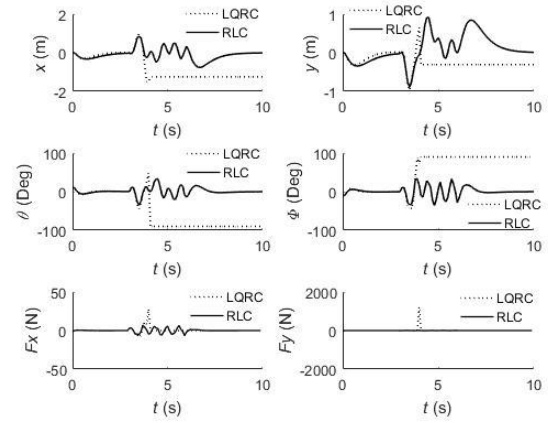
کنترل پاندول معکوس چهار درجه آزادی استفاده شده است. معادلات حاکم بر سیستم با استفاده از روش لاگرانژ استخراج شده اند. ابتدا برای اعتبار سنجی کنترلر LQR با استفاده از روش یادگیری تقویتی، کنترلر برای پاندول معکوس دو درجه آزادی طراحی و برای مقدار زاویه انحراف و جابجایی اولیه مختلف آزمایش و نتایج با کنترلر LQR مقایسه شده است، نتایج نشان دهنده این می‌باشد که LQR با استفاده از روش یادگیری تقویتی، سرعت کنترل و محدوده تغییر پارامتر بیشتری دارد. در ادامه کنترلر LQR با

استفاده از روش یادگیری تقویتی برای پاندول چهار درجه آزادی طراحی و در شرایط مختلف آزمایش و نتایج با کنترلر LQR مقایسه شده است. شرایط آزمایش عبارت اند از: 1- انحراف زاویه اولیه و جابجایی اولیه پایه پاندول بزرگ 2- نامعینی های پارامتری و ساختاری 3- اغتشاش وارد شده توسط نیروی ضربه ای و غیر ضربه ای وارد به پایه و جرم متمرکز 4- نویز سنسورها. پس از شبیه سازی نتایج، مشاهده شد که کنترلر LQR با استفاده از روش یادگیری تقویتی نسبت به کنترلر LQR در برابر انحرافات اولیه بزرگ، نامعینی های ساختاری و پارامتری، اغتشاشات وارده و نویز سنسورها عملکرد خوبی داشته است.



a) The response of the system to the non-impact force in the x and y directions
$L = 35\ \text{cm}, Ml = 150\ \text{g}, M = 100\ \text{g}$
$x = 0\ \text{m}, y = 0\ \text{m}, \theta = 10^{\circ}, \varphi = -10^{\circ}$
$t = 3 - 6\ \text{s}$ در $\overline{F_x} = 3\sin t\ \text{N}$
$t = 3 - 6\ \text{s}$ در $\overline{F_y} = 3\sin t\ \text{N}$

الف) پاسخ سیستم نسبت به نیروی غیر ضربه ای در راستاهای x و y

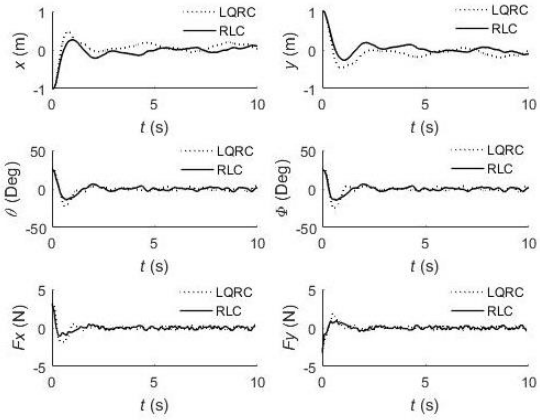


b) The response of the system to the non-impact force in the θ and φ directions
$L = 35\ \text{cm}, Ml = 150\ \text{g}, M = 100\ \text{g}$
$x = 0\ \text{m}, y = 0\ \text{m}, \theta = 10^{\circ}, \varphi = -10^{\circ}$
$t = 3 - 6\ \text{s}$ در $\overline{F_\theta} = 0.45\sin t\ \text{N}$
$t = 3 - 6\ \text{s}$ در $\overline{F_\varphi} = 0.45\sin t\ \text{N}$

ب) پاسخ سیستم نسبت به نیروی غیر ضربه ای در راستاهای θ و φ

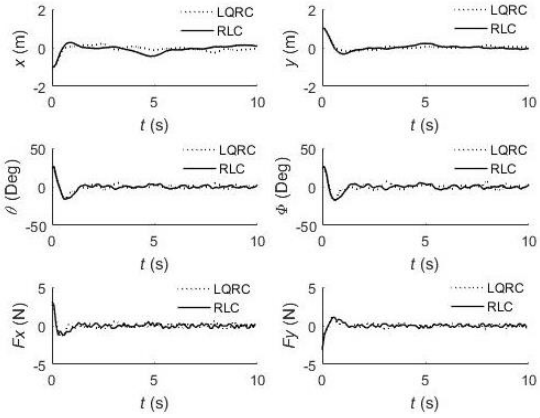
Fig.7 Time response of Reinforcement learning controller and LQR controller for four-degree of freedom inverted pendulum for continuous disturbance

شکل 7 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول چهار درجه آزادی برای اغتشاش پیوسته



a) System response to noise in the amount of angles received
$L = 35\ \text{cm}, Ml = 150\ \text{g}, M = 100\ \text{g}$
$x = -1\ \text{m}, y = 1\ \text{m}, \theta = 25^{\circ}, \varphi = 25^{\circ}$

الف) پاسخ سیستم نسبت به نویز در مقدار زاویه دریافتی



b) System response to noise in the all amount received
$L = 35\ \text{cm}, Ml = 150\ \text{g}, M = 100\ \text{g}$
$x = -1\ \text{m}, y = 1\ \text{m}, \theta = 25^{\circ}, \varphi = 25^{\circ}$

ب) پاسخ سیستم نسبت به نویز در تمامی مقادیر دریافتی

Fig.8 Time response of Reinforcement learning controller and LQR controller for four-degree of freedom inverted pendulum for sensor noises

شکل 8 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول چهار درجه آزادی برای نویز سنسورها

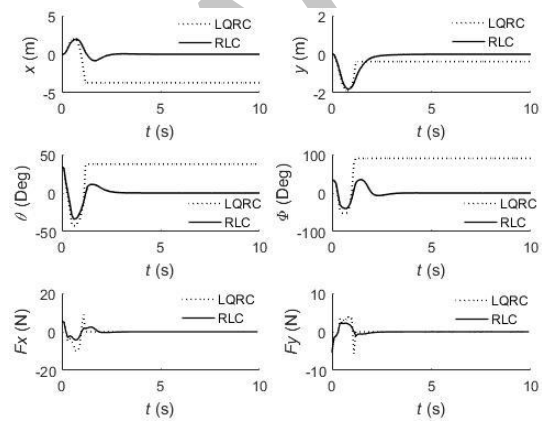


Fig.9 Time response of Reinforcement learning controller and LQR controller for four-degree of freedom inverted pendulum for big angles
$L = 35\ \text{cm}, Ml = 150\ \text{g}, M = 100\ \text{g}$
$x = 0\ \text{m}, y = 0\ \text{m}, \theta = 33^{\circ}, \varphi = 33^{\circ}$

شکل 9 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول چهار درجه آزادی برای زوایای بزرگ

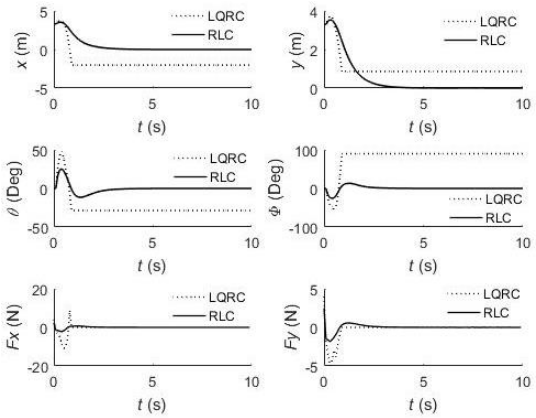


Fig.10 Time response of Reinforcement learning controller and LQR controller for four-degree of freedom inverted pendulum for big position
$L = 35$ cm , $Ml = 150$ g , $M = 100$ g
$x = 3.3$ m , $y = 3.3$ m , $\theta = 0^{\circ}$, $\varphi = 0^{\circ}$

شکل 10 پاسخ زمانی کنترلر یادگیری تقویتی و LQR برای پاندول چهار درجه آزادی برای جابجایی بزرگ پایه پاندول

5- مراجع

[1] Y. Li, L. Chen, K. P. Tee, Q. Li, Reinforcement learning control for coordinated manipulation of multi-robots, *Neurocomputing*, Vol. 170, No. 1, pp. 168-175, 2015.
[2] F. Farivar, M. N. Ahmadabadi, Continuous reinforcement learning to robust fault tolerant control for a class of unknown nonlinear systems, *Applied Soft Computing*, Vol. 37, No. 1, pp. 702-714, 2015.
[3] B. Fernandez-Gauna, M. Graña, J. M. Lopez-Guede, I. Etxeberria-Agiriano, I. Ansoategui, Reinforcement learning endowed with safe veto policies to learn the control of Linked-Multicomponent robotic systems, *Information Sciences*, Vol. 317, No. 1, pp. 25-47, 2015.
[4] M. I. El-Hawwary, A. L. Elshafei, H. M. Emara, H. A. Abdel Fattah, Adaptive fuzzy control of the inverted pendulum problem, *IEEE Transaction on Control Systems Technology*, Vol. 14, No. 6, pp. 1135-1144, 2006.
[5] J. J. Wang, Stabilization and tracking control of X–Z inverted pendulum with sliding-mode control, *ISA Transaction*, Vol. 51, No. 6, pp. 763–770, 2012.
[6] R. M. Brisilla, V. Sankaranarayanan, Nonlinear control of mobile inverted pendulum, *Robotics and Autonomous Systems*, Vol. 70, No. 1, pp. 145–155, 2015.
[7] V. Casanova, J. Alcaína, J. Salt, R. Pizá, Á. Cuenca, Control of the rotary inverted pendulum through threshold-based communication control, *ISA Transactions*, Vol. 62, No. 1, pp. 357–366, 2016.
[8] A. Kharola, P. Patil, S. Raiwani, D. Rajput, A comparison study for control and stabilisation of inverted pendulum on inclined surface (IPIS) using PID and fuzzy controllers, *Perspectives in Science*, Vol. 8, No. 1, pp. 187–190, 2016.
[9] A. I. Roose, S. Yahya, H. Al-Rizzo, Fuzzy-logic control of an inverted pendulum on a cart, *Computers and Electrical Engineering*, Vol. 61, No. 1, pp. 31–47, 2017.
[10] J. Ghommama, F. Mnif, Predictor-based control for an inverted pendulum subject to networked time delay, *ISA Transactions*, Vol. 67, No. 1, pp. 306–316, 2017.
[11] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, pp. 70-140, Cambridge-Massachusetts-London-England, The MIT Press, 2005.