مجله کنترل
ISSN 2008-8345
جلد ۸، شماره ۱، بهار ۱۳۹۳، صفحه ۲۰-۱۱

یک روش ترکیبی جدید یادگیری تقویتی فازی

فرزانه قربانی[۱]، ولی درهمی[۲]، حسین نظام آبادی پور[۳]

[۱] فارغ‌التحصیل کارشناسی ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه یزد، f.ghorbani@stu.yazd.ac.ir
[۲] دانشیار، دانشکدهٔ مهندسی برق و کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه یزد، vderhami@yazd.ac.ir
[۳] استاد، دانشکدهٔ فنی و مهندسی، گروه مهندسی برق، دانشگاه شهید باهنر کرمان، nezam@uk.ac.ir

(تاریخ دریافت مقاله ۱۳۹۳/۱/۱۳، تاریخ پذیرش مقاله ۱۳۹۳/۳/۸)

چکیده: در این مقاله یک روش جدید یادگیری تقویتی پیوسته برای مسائل کنترل ارائه می‌شود. روش ارائه شده از ترکیب روش "تکرار سیاست کمترین مربعات " با یک سیستم فازی سوگنوی مرتبه صفر حاصل شده و "تکرار سیاست کمترین مربعات فازی" نامیده شده است. در اینجا برای هر قاعده فازی تعدادی عمل نامزد در نظر گرفته می‌شود. هدف، یافتن مناسب‌ترین عمل نامزد (تالی) برای هر قاعده می‌باشد. با استفاده از بردار شدت آتش قواعد فازی و عمل‌های نامزد مربوط به قواعد، توابع پایه حالت –عمل به گونه‌ای تعریف شده‌اند که شرایط قضایای روش تکرار سیاست کمترین مربعات را برآورده می‌نمایند. با استفاده از توابع پایه حالت- عمل تعریف شده و بهره‌گیری از الگوریتم تکرار سیاست کمترین مربعات، یک روش جدید برای تازه‌سازی پارامترهای وزن تالی قواعد ارائه می‌شود. تحلیل ریاضی که برای این الگوریتم آورده می‌شود، کران خطایی برای اختلاف تابع مقدار ارزش حالت-عمل واقعی و تخمین تابع ارزش حالت-عمل حاصل از الگوریتم ارائه شده، تعریف می‌کند. نتایج شبیه‌سازی در مساله معروف قایق، حاکی از سرعت آموزش بالاتر و نیز کیفیت عملکرد بهترِ روش پیشنهادی نسبت به دو روش یادگیری کیوی فازی و یادگیری سارسای فازی است. از مزایای دیگر روش ارائه شده، عدم نیاز به تعیین نرخ آموزش است.

کلمات کلیدی: یادگیری تقویتی، تکرار سیاست کمترین مربعات ، تقریب تابع ارزش حالت-عمل، سیستم فازی.

A Novel approach in Fuzzy Reinforcement Learning

Farzaneh Ghorbani,Vali Derhami, Hossein Nezamabadipour

Abstract: In this paper, we present a novel continuous reinforcement learning approach. The proposed approach, called "Fuzzy Least Squares Policy Iteration (FLSPI)", is obtained from combination of "Least Squares Policy Iteration (LSPI)" and a zero order Takagi Sugeno fuzzy system. We define state-action basis function based on fuzzy system so that LSPI conditions are satisfied. It is proven that there is an error bound for difference of the exact state-action value function and approximated state-action value function obtained by FLSPI. Simulation results show that learning speed and operation quality for FLSPI are higher than two previous critic-only fuzzy reinforcement learning approaches i.e. fuzzy Q-learning and fuzzy Sarsa learning. Another advantage of this approach is needlessness to learning rate determination.

Keywords: Reinforcement learning; least squares policy iteration, state-action function approximation, fuzzy system.

۱- مقدمه

با توجه به ویژگی‌های مطلوب یادگیری تقویتی مانند عدم نیاز به داده‌های آموزشی در مسائلی که داده‌های آموزشی در دسترس نیست و همچنین آموزش، تنها با استفاده از یک سیگنال اسکالر، این روش‌ها مورد توجه هستند. روش‌های یادگیری تقویتی، قدرتمند هستند، اما روش‌های یادگیری تقویتی استاندارد بر روی فضای حالت و عمل گسسته، عمل می‌کنند. این روش‌ها در مسائل با فضای بزرگ یا فضای پیوسته دچار چالش می‌شوند، زیرا گسسته سازی این فضاها مشکل است و آموزش، نیاز به زمان و فضای زیاد دارد. این مشکل در فضاهای بزرگ، نفرین یا تنگنای ابعاد[1] نام دارد. برای رفع مشکل تنگنای ابعاد، از ترکیب تقریب زننده های تابع با روش‌های یادگیری تقویتی استفاده شده است. این روش‌ها از ترکیب تقریب زننده‌هایی مانند سیستم های فازی و شبکه‌های عصبی با روش‌های یادگیری تقویتی به وجود می‌آیند. استفاده از تقریب زننده‌ها در زمره کاربردهای موفق روش‌های یادگیری تقویتی قرار دارند. اول به دلیل این که نیاز به نگه‌داری جدولی اطلاعات تابع ارزش حالت-عمل و لذا حافظه زیاد ندارند. دوم این که نیاز به اطلاعات دقیق محیط ندارند. در ضمن استفاده از آن‌ها ساده است و عملکردشان واضح و شفاف می‌باشد، لذا از لحاظ تحلیل و اشکال زدایی ساده می‌باشند [۱].

تمرکز ما در این مقاله بر روی روش‌های یادگیری تقویتی فازی با معماری نقاد-تنها است. الگوریتم‌های یادگیری کیوی فازی [۲-۷] و یادگیری سارسای فازی [۸] از جمله الگوریتم‌های ارائه شده در یادگیری تقویتی فازی با معماری نقاد-تنها هستند. این روش‌ها از لحاظ ارائه تحلیل ریاضی مثبت در فضاهای پیوسته و وابستگی به نرخ آموزش دچار چالش هستند. ما در این مقاله به دنبال ارائه روشی هستیم که چالش‌های مذکور را برطرف نموده و نسبت به روش‌های قبل، دارای کارایی بهتری باشد.

یکی از روش‌های یادگیری تقویتی که دارای تحلیل ریاضی مثبت، عدم وابستگی به نرخ آموزش و کارایی مناسب است، تکرار سیاست کمترین مربعات (LSPI[2]) [۱،۹،۱۰] می‌باشد. برای پیاده سازی و استفاده عملی از LSPI، تعریف توابع پایه ضروری است اما در مقاله‌های مذکور، شیوه تعریف توابع پایه مشخص نشده است. LSPI یک الگوریتم تکراری است که در هر تکرار دارای دو مرحله است: (الف) ارزیابی سیاست، مقدار تابع ارزش حالت-عمل را برای سیاست جاری محاسبه می‌کند و (ب) بهبود سیاست، با استفاده از تابع ارزش حالت-عمل مرحله قبل سیاست جدید را تعریف می‌کند. روش LSPI بر روی فضای حالت و عمل متناهی ارائه شده است. تحلیل ریاضی LSPI با استناد به تحلیل ریاضی تکرار سیاست (PI[3]) [۱۱] بیان شده است. در LSPI نحوه تعریف دقیق توابع پایه بیان نشده است. مرجع [۱۲]، این روش را به صورت LSPI برخط[4]، برای فضای حالت نامتناهی اصلاح کرده است، اما در این روش نیز توابع پایه به صورت دقیق بیان نشده است و هیچ تحلیل ریاضی و تضمین همگرایی برای این روش (در فضای نامتناهی) وجود ندارد.

در این مقاله از سیستم های فازی به عنوان یک تقریب زننده جامع [۱۳-۱۵] برای تقریب تابع ارزش حالت-عمل در روش LSPI استفاده شده و منجر به تعریف روش تکرار سیاست کمترین مربعات فازی (FLSPI[5]) شده است. در واقع از سیستم های فازی برای تعریف توابع پایه در LSPI استفاده شده است. از سوی دیگر، سیستم فازی مورد استفاده در FLSPI، سیستمی است که نیاز به تنظیم تالی‌های قواعد دارد و می‌توان گفت از LSPI برای تنظیم تالی های این سیستم فازی استفاده شده است.

آن چه که در ادامه این مقاله خواهد آمد به این صورت بخش بندی شده است: در بخش دوم مروری کوتاه بر یادگیری تقویتی به طور کلی و سپس الگوریتم LSPI به عنوان یکی از روش‌های مورد استفاده در یادگیری تقویتی انجام شده است. در بخش سوم به روش FLSPI که برای اولین بار در این مقاله ارائه شده پرداخته می‌شود. بخش چهارم در بردارنده نتایج شبیه سازی مساله قایق با استفاده از FLSPI است و در بخش پنجم، جمع بندی و نتایج حاصل از به کارگیری این الگوریتم آمده است.

۲- یادگیری تقویتی

در یک سیستم مبتنی بر عامل با یادگیری تقویتی، در هر قدم زمانی t، عامل، حالت فعلی s_t از فضای حالت S مشاهده نموده و عملی را از فضای عمل متناهی A، بر اساس سیاست π انتخاب و به محیط اعمال می‌کند. در پی آن، محیط به حالت جدید s_{t+1} از فضای حالت S با احتمال انتقال $\mathcal{P}(s_t, a_t, s_{t+1})$ رفته و عامل سیگنال تقویتی $\mathcal{R}_{t+1} = \mathcal{R}(s_t, a_t)$ را دریافت می‌کند [۶۱].

سیاست که با $\pi(s, a)$ نشان داده می‌شود، احتمال انتخاب عمل a در حالت s می‌باشد. مبنای کار در یادگیری تقویتی بر اساس پاداش و جریمه است و هدف، حداکثر کردن مجموع پاداش‌های دریافتی در طول یادگیری می‌باشد. بر این اساس عامل یاد می‌گیرد عملی را انتخاب کند که او را به حالتی با بیش‌ترین ارزش برساند. ارزش حالت s تحت سیاست π توسط تابع زیر تعریف می‌شود [۶۱]:

$$V^{\pi}(s) = E_{\pi}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_t | s_t = s] \quad 0 \le \gamma \le 1 \qquad (۱)$$

1 Curse of dimensionality
2 Least Squares Policy Iteration (LSPI)
3 Policy Iteration (PI)
4 Online
5 Fuzzy Least Squares Policy Iteration

که γ نرخ کاهش و $\mathrm{E}_{\pi}[.]$ امید ریاضی می‌باشد. به طور مشابه، تابع ارزش حالت-عمل برای عمل a و حالت s که مشخص می کند در حالت s ارزش انجام عمل a چه اندازه است، به صورت زیر تعریف می‌شود [۶۱]:

$$Q^{\pi}(s,a) = E_{\pi}[\sum_{t=0}^{\infty} \gamma^{t} \mathcal{R}_{t} | s_{t} = s, a_{t} = a] \quad (۲)$$

۲-۱-تکرار سیاست کمترین مربعات

روش تکرار سیاست (PI[1]) [۷۱] یکی از روش‌های یادگیری تقویتی است که سیاست بهینه را برای هر زنجیره تصمیم مارکوف به وسیله تولید دنباله ای از سیاست‌ها در فضای گسسته به دست می‌آورد. PI، یک الگوریتم تکراری است که در تکرار mام آن دو مرحله وجود دارد: ارزیابی سیاست، مقدار تابع Q^{π_m} را از سیاست فعلی π_m محاسبه می کند:

$$\begin{aligned} Q^{\pi_m}(s,a) &= \mathcal{R}(s,a) \\ &+\gamma \sum_{s' \in S} \mathcal{P}(s,a,s') \sum_{a' \in A} \pi(s',a') Q^{\pi_m}(s',a') \end{aligned} \quad (۳)$$

بهبود سیاست، سیاست حریصانه بهبود یافته π_{m+1} را روی Q^{π_m} تعریف می کند:

$$\pi_{m+1}(s) = \underset{a \in A}{argmax}\, Q^{\pi_m}(s,a) \quad (۴)$$

این روش در فضای حالت و عمل متناهی دارای کارایی بالایی است و کران خطایی برای آن تعریف شده است [۱۱].

مقدار دقیق Q^{π_m} را می‌توان از حل معادله زیر به دست آورد:

$$T_{\pi_m} Q = Q \quad (۵)$$

که T_{π} عملگر بلمن تحت سیاست π می‌باشد و به صورت زیر تعریف می‌شود [۱]:

$$\begin{aligned} (T_{\pi}Q)(s,a) &= \mathcal{R}(s,a) \\ &+\gamma \sum_{s' \in S} \mathcal{P}(s,a,s') \sum_{a' \in A} \pi(s',a') Q(s',a') \end{aligned} \quad (۶)$$

این روش مانند روش‌های دیگر یادگیری تقویتی در فضاهای بزرگ، دچار مشکل تنگنای ابعاد می‌باشد.

مرجع [۱] روش تکرار سیاست کمترین مربعات را با استفاده از روش تکرار سیاست برای رفع مشکل تنگنای ابعاد، ارائه نموده است. این روش از روش‌هایی است که به جای محاسبه دقیق تابع ارزش حالت-عمل Q، تقریب آن یعنی $\hat{Q}$ را محاسبه می‌نماید. مبنای LSPI، از بین بردن خطای حاصل از $\hat{Q}$ و تصویر[2] آن تحت عملگر بلمن است.

در حالت کلی می‌توان از یک ترکیب خطی وزن دار برای تقریب تابع ارزش حالت-عمل به صورت زیر استفاده کرد:

$$\hat{Q}^{\pi} = \boldsymbol{\Phi W} \quad (۷)$$

که در آن $\boldsymbol{W}$ ماتریس وزن‌ها و $\boldsymbol{\Phi}$ ماتریس توابع پایه حالت- عمل هستند:

$$\boldsymbol{\Phi} = \begin{pmatrix} \emptyset(s_1,a_1)^T \\ \dots \\ \emptyset(s_i,a_i)^T \\ \dots \\ \emptyset(s_{|S|},a_{|A|})^T \end{pmatrix} \quad (۸)$$

به طوری که

$$\emptyset(s,a) = \begin{pmatrix} \emptyset_1(s,a) \\ \dots \\ \emptyset_k(s,a) \end{pmatrix}. \quad (۹)$$

که $k \ll |S||A|$ تعداد توابع پایه، $\emptyset_j$ توابع پایه حالت – عمل می باشند.

در [۱] LSPI پس از انجام محاسبات نتیجه گرفته شده که می‌توان $\hat{Q}^{\pi}$ را از حل معادله (۱۰) به دست آورد:

$$\hat{Q}^{\pi} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi} T_{\pi} \hat{Q}^{\pi} \quad (۱۰)$$

از ساده سازی رابطه بالا (پس از انجام محابات مبسوط که مجال بیان آن‌ها در این مقاله نیست)، می‌توان به نتایج زیر رسید [۱]:

$$AW = b \quad (۱۱)$$

$$\begin{aligned} A_{new} = A_{old} + \emptyset(s,a) \Big(&\emptyset(s,a) \\ &- \gamma \emptyset\big(s', \pi(s')\big) \Big)^T \end{aligned} \quad (۱۲)$$

$$b_{new} = b_{old} + \emptyset(s,a)\mathcal{R} \quad (۱۳)$$

که $\mathcal{R}$ پاداش انتقال از حالت s به حالت s' است وقتی که عمل a انتخاب شده باشد. γ، نرخ کاهش است و ماتریس A و بردار b که به صورت تکراری محاسبه می‌شوند، برای محاسبه بردار وزن W مورد استفاده قرار می گیرند.

1 Policy Iteration

2 Projection

۳- تکرار سیاست کمترین مربعات فازی(FLSPI)

در این بخش با استفاده از سیستم‌های فازی به عنوان تقریب زننده روش تکرار سیاست کمترین مربعات فازی (FLSPI) ارائه می‌شود. FLSPI، از سیستم فازی سوگنو مرتبه صفر استفاده می‌نماید. R قاعده سیستم به شکل زیر تعریف می‌شود:

$$R_i: If\ x_1\ is\ L_{i1}\ and\ \dots and\ If\ x_n\ is\ L_{in}, Then$$

$$(o_{i1}\ with\ weigth\ w^{i1}\ or\ \dots or\ o_{im}\ with\ weigth\ w^{im})$$

که $s = (x_1, \dots, x_n)$ برداری از فضای ورودی n-بعدی و $L_i = L_{i1} \times \dots \times L_{in}$، فضای n-بعدی محدب است و مجموعه‌های فازی قاعده iام دارای مراکز یکتا هستند. m، تعداد عمل‌های ممکن برای هر قاعده و o_{ij}، jامین عمل نامزد برای قاعده iام با وزن w^{ij} می‌باشد. مشخصات توابع عضویت، تعداد عملها و مقدار عملها وابسته به مساله بوده و با توجه به تجربه طراح تعیین می شوند.

برای قاعده iام با توجه به وزن هر عمل و استفاده از روش انتخاب عمل شبه حریصانه، عمل o_{ii^+} (i^+ اندیس عمل مورد نظر است) انتخاب می‌شود و عمل نهایی به صورت زیر محاسبه می‌شود:

$$a_t(s_t) = \sum_{i=1}^{R} \mu_i(s_t) o_{ii^+} \qquad (۱۴)$$

که $\mu_i(s)$، شدت آتش نرمال شده قاعده iام به ازای ورودی s است.

هدف از آموزش سیستم تقریب w^{ij} به صورت برخط، برای رسیدن به سیاست بهینه است.

شدت آتش هر قاعده از حاصل ضرب مقدم‌های مجموعه‌های فازی به دست می‌آید و با استفاده از توابع شدت آتش نرمال شده قواعد، توابع پایه حالت-عمل را به صورت زیر تعریف می‌نماییم:

$$\varphi(s,a) = \left[\overbrace{0\dots\mu_1(s)\dots0}^{m}\ \overbrace{0\dots\mu_2(s)\dots0}^{m}\ \dots\ \overbrace{0\dots\mu_R(s)\dots0}^{m}\right]^T \qquad (۱۵)$$

که در آن، m تعداد عمل‌ها می‌باشد. لازم به ذکر است تعدادعملهای نامزد و این پارامتر توسط کاربر، بر اساس مساله و تجربه کاربر تعیین می‌شود.

روال الگوریتم FLSPI را می‌توان به صورت زیر خلاصه کرد:

۱-حالت اولیه s_0 را مشاهده کن.

۲-در مجموعه عمل‌های نامزد هر قاعده، یکی از عمل‌ها را با توجه به مقدار وزنشان و روش انتخاب عمل شبه حریصانه[1] انتخاب کن.

۳-تا زمانی که به انتهای اپیزود نرسیده ای مراحل زیر را تکرار کن:

a. با استفاده از رابطه (۱۴) عمل نهایی a_t را محاسبه کن.

b. با اعمال a_t به محیط، حالت s_{t+1} را مشاهده کن و پاداش r_{t+1} را دریافت کن.

c. ماتریس A را با استفاده از رابطه زیر به روزرسانی کن:

(۱۶)

$$A_{t+1} = A_t + \emptyset(s_t,a_t)\left(\emptyset(s_t,a_t) - \gamma\emptyset\left(s_{t+1},\pi(s_{t+1})\right)\right)^T$$

d. بردار b را با استفاده از رابطه زیر به روزرسانی کن:

$$b_{t+1} = b_t + \emptyset(s_t,a_t) r_{t+1} \qquad (۱۷)$$

e. $t \leftarrow t+1$

۴-رابطه

$$\frac{1}{t-1} A_t W_k = \frac{1}{t-1} b_t \qquad (۱۸)$$

را برای به دست آوردن W_k حل کن و $k \leftarrow k+1$.

تا زمانی که شرط توقف برقرار نیست، قرار بده $s_t = s_0$ و مراحل ۲ تا ۴ را تکرار کن.

که شرط توقف یا رسیدن به هدف (با خطای تعریف شده توسط کاربر) و یا تعداد تکرار خاصی که توسط کاربر تعریف می‌شود، است.

به روز رسانی ماتریس A و بردار b، توسط روابط (۱۶) و (۱۷)، در هر قدم زمانی و به روزرسانی بردار وزن W توسط رابطه (۱۸)، در هر اپیزود انجام می‌شود. پارامترهای A و b، همان پارامترهای مطرح شده در روابط (۱۳) تا (۱۵) هستند که برای محاسبه بردار وزن W مورد استفاده قرار می‌گیرند. به عبارت دیگر، سیاست در طول یک اپیزود، تغییر نمی‌کند. در صورتی که این الگوریتم برای موارد غیر اپیزودیک استفاده شود، به جای به روز رسانی بردار وزن بعد از هر اپیزود، این بردار، بعد از تعداد مشخصی قدم زمانی به روز رسانی می‌شود. شرط توقف رسیدن به هدف با یک خطای قابل قبول یا تعداد تکرار مشخصی می‌باشد.

لازم به یادآوری است که به روز رسانی بردار W در واقع به روز رسانی $\hat{Q}$ می‌باشد و منجر به استخراج سیاست جدید می‌شود.

٤- تحلیل ریاضی روش پیشنهادی

در این بخش، مباحثی مطرح می‌شود که منجر به تعریف کران برای خطای تقریب تابع حالت-عمل و مقدار بهینه تابع حالت-عمل می‌شود. با استفاده از مرجع [۱۳]، قضیه زیر را داریم.

تبصره ۱ (قضیه استون-وایرشتراس): اگر Z یک مجموعه از توابع پیوسته روی فضای محدب X باشد و اگر

الف) Z یک جبر باشد یعنی نسبت به جمع و ضرب اسکالر و ضرب بسته باشد.

1 ε-greedy

ب) Z نقاط روی X را جدا کند:

$$\forall x_1, x_2 \in X, x_1 \neq x_2;\ \exists F \in Z\ s.t\ F(x_1) \neq F(x_2)$$

ج) Z روی هیچ نقطه ای از X صفر نشود:

$$\forall x \in X;\ \exists F \in Z\ s.t\ F(x) \neq 0$$

آنگاه برای هر تابع پیوسته G(x) روی X و هر ε>0 دلخواه، یک $F \in Z$ وجود دارد به‌طوری‌که

$$\|F(x) - G(x)\|_\infty < \varepsilon$$

لم ۱- برای هر تابع پیوسته Q و هر ε>0 دلخواه، تابع $\hat{Q} = \sum_{i=1}^{R} \mu_i(s) w_{ii^+}$ وجود دارد به‌طوری‌که

$$\left\|\hat{Q} - Q\right\|_\infty < \varepsilon$$

اثبات- X را فضای حاصل از ضرب دکارتی فضای حالت (زیرفضایی محدب از $\mathbb{R}^n$) و فضای عمل (زیرفضایی محدب از $\mathbb{R}^m$) در نظر بگیرید، در این صورت X زیرفضایی محدب از $\mathbb{R}^{n+m}$ است که n بعد فضای حالت و m بعد فضای عمل می‌باشد. همچنین Q روی این فضا تابعی پیوسته است. اکنون با تعریف Z به عنوان فضای همه توابع $\hat{Q} = \boldsymbol{\Phi} W$ که W یک بردار است و $\boldsymbol{\Phi}$ با رابطه (۸) مشخص می‌شود، کافی است شرایط قضیه۱ ارضا شود تا اثبات کامل شود:

تعریف می‌کنیم:

$$\hat{Q}: X \to Y\ , y = \hat{Q}(s,a)\ , X \subseteq \mathbb{R}^{n+m}\ , Y \subseteq \mathbb{R}$$

همچنین R قاعده به به صورت تعریف شده در بخش ۳، در نظر می‌گیریم:

$$R_i: If\ s_1\ is\ L_{i1}\ and\ \dots and\ If\ s_n\ is\ L_{in}, Then$$

$$(o_{i1}\ with\ weigth\ w^{i1}\ or\ \dots or\ o_{im}\ with\ weigth\ w^{im})$$

اکنون y را به این شکل تعریف می‌کنیم:

$$y = \hat{Q}(s,a) = \sum_{i=1}^{R} \mu_i(s) W_{ii^+} \qquad (۱۹)$$

که در آن W_{ii^+} وزن مربوط به عمل انتخاب شده در قاعده i ام و $\mu_i(s)$ شدت آتش نرمال شده قاعده i ام به ازای ورودی s است:

$$\mu_i(s) = \frac{\alpha_i(s)}{\sum_{j=1}^{R} \alpha_j(s)} \qquad (۲۰)$$

که $\alpha_i(s)$ شدت آتش قاعده i ام به ازای ورودی s می باشد:

$$\alpha_i(s) = \prod_{l=1}^{n} \mu_i^l(s_l) \qquad (۲۱)$$

$\mu_i^l(s_l)$ درجه عضویت s_l در l امین تابع عضویت تعریف شده می‌باشد. لذا داریم:

$$\hat{Q}(s,a) = \frac{\sum_{i=1}^{R} W_{ii^+} \prod_{l=1}^{n} \mu_i^l(s_l)}{\sum_{i=1}^{R} \prod_{l=1}^{n} \mu_i^l(s_l)} \qquad (۲۲)$$

اکنون با استفاده از دو رابطه زیر شرط قسمت (الف)تبصره ۱ را بررسی می کنیم:

$$\left(\sum_i p_i b_i\right)\left(\sum_j c_j\right) + \left(\sum_j q_j c_j\right)\left(\sum_i b_i\right) = \sum_i \sum_j (p_i + q_j)(b_i c_j) \qquad (۲۳)$$

$$\left(\sum_i a_i\right)\left(\sum_j b_j\right) = \sum_i \sum_j a_i b_j \qquad (۲۴)$$

دو تابع دلخواه $\hat{Q}_1$ با R_1 قاعده و $\hat{Q}_2$ با R_2 قاعده از Z را در نظر بگیرید:

$$\hat{Q}_1(s,a) = \frac{\sum_{i=1}^{R_1} W_{ii^+} \prod_{l=1}^{n} \mu_i^l(s_l)}{\sum_{i=1}^{R_1} \prod_{l=1}^{n} \mu_i^l(s_l)}$$

$$\hat{Q}_2(s,a) = \frac{\sum_{j=1}^{R_2} W_{jj^+} \prod_{l=1}^{n} \mu_j^l(s_l)}{\sum_{j=1}^{R_2} \prod_{l=1}^{n} \mu_j^l(s_l)}$$

با استفاده از رابطه (۲۳) داریم:

$$\hat{Q}_1(s,a) + \hat{Q}_2(s,a) = \frac{\sum_{i=1}^{R_1}\sum_{j=1}^{R_2}(W_{ii^+} + W_{jj^+})(\prod_{l=1}^{n} \mu_i^l(s_l)\mu_j^l(s_l))}{\sum_{i=1}^{R_1}\sum_{j=1}^{R_2}(\prod_{l=1}^{n} \mu_i^l(s_l)\mu_j^l(s_l))}$$

در نتیجه $\hat{Q}_1(s,a) + \hat{Q}_2(s,a) \in Z$. پس Z نسبت به جمع بسته است.

حال با استفاده از رابطه (۲۴) خواهیم داشت:

$$\hat{Q}_1(s,a).\hat{Q}_2(s,a) = \frac{\sum_{i=1}^{R_1}\sum_{j=1}^{R_2}(W_{ii^+}.W_{jj^+})(\prod_{l=1}^{n} \mu_i^l(s_l)\mu_j^l(s_l))}{\sum_{i=1}^{R_1}\sum_{j=1}^{R_2}(\prod_{l=1}^{n} \mu_i^l(s_l)\mu_j^l(s_l))}$$

و در نتیجه $\hat{Q}_1(s,a).\hat{Q}_2(s,a) \in Z$ یعنی Z نسبت به ضرب بسته است. به وضوح Z نسبت به ضرب اسکالر هم بسته است. لذا Z یک جبر است.

برای بررسی قسمت (ب) تبصره ۱، فرض کنید

$$v = (s_1, a_1), h = (s_2, a_2) \in X$$

دو بردار دلخواه باشند که $v \neq h$، سیستم فازی با دو قاعده را در نظر بگیرید که دارای دو تابع عضویت به شکل زیر است:

$$\mu_1^l(x_l) = \exp(-\frac{1}{2}(x_l - v_l)^2)$$

$$\mu_2^l(x_l) = \exp(-\frac{1}{2}(x_l - h_l)^2)$$

و همچنین تالی‌های:

$$o_{11} = a_1/2\ , o_{12} = (a_2/2)\exp\left(\frac{1}{2}\|v-h\|^2\right)$$

$$o_{21} = (a_1/2)\exp\left(\frac{1}{2}\|v-h\|^2\right), o_{22} = a_2/2$$

با وزن‌های:

$$W_{11} = W_{12} = 0\,, W_{21} = W_{22} = 1$$

که اندیس‌های منتخب برای a_1، قاعده اول اندیس اول، قاعده دوم اندیس اول و برای a_2، قاعده اول اندیس دوم و قاعده دوم اندیس دوم می‌باشد.

شدت آتش‌ها به صورت زیر به دست می‌آیند:

$$\alpha_1(x) = \prod_{l=1}^{n} \mu_1^l(s_l) = \exp(-\frac{1}{2}\|x - v\|_2^2)$$

$$\alpha_2(x) = \prod_{l=1}^{n} \mu_2^l(s_l) = \exp(-\frac{1}{2}\|x - h\|_2^2)$$

لذا داریم:

$$\hat{Q}(s_1, a_1) = \frac{\exp(-\frac{1}{2}\|v - h\|_2^2)}{1 + \exp(-\frac{1}{2}\|v - h\|_2^2)}$$

$$\hat{Q}(s_2, a_2) = \frac{1}{1 + exp(-\frac{1}{2}\|v - h\|_2^2)}$$

در نتیجه $\hat{Q}(v) \neq \hat{Q}(h)$. پس Z نقاط روی X را جدا می‌کند.

برای بررسی قسمت (ج) کافی است با تعاریف قسمت قبل $W_{ij} > 0$ برای همه i و jها. در این صورت به ازای هر x، $\hat{Q}(x) \neq 0$. پس Z روی هیچ نقطه ای از X صفر نمی‌شود. ■

آنچه از لم۱ نتیجه می‌شود این است که برای هر تابع دلخواه می‌توان تابعی از سیستم فازی استفاده شده در الگوریتم FLSPI یافت که با هر دقت دلخواهی آن را تقریب بزند.

لم ۲- برای سیاست ایستای π و اسکالر دلخواه x داریم:

$$T(Q + xe)(s,a) = TQ(s,a) + \gamma x \quad (۲۵)$$

$$T_\pi(Q + xe)(s,a) = T_\pi Q(s,a) + \gamma x \quad (۲۶)$$

که در آن e همانی است.

اثبات- چون این مساله یک مساله کاهشی[1] با نرخ کاهش γ است، با استناد به مرجع [۱۱] اثبات کامل می‌شود. ■

لم ۳- برای سیاست ایستای π داریم:

$$\|TQ - T\hat{Q}\|_\infty \leq \gamma\|Q - \hat{Q}\|_\infty \quad (۲۷)$$

$$\|T_\pi Q - T_\pi\hat{Q}\|_\infty \leq \gamma\|Q - \hat{Q}\|_\infty \quad (۲۸)$$

اثبات- چون این مساله یک مساله کاهشی با نرخ کاهش γ است، با استناد به مرجع [۱۱] اثبات کامل می‌شود. ■

لم ٤- فرض کنید π_k سیاست تولید شده توسط kامین تکرار از الگوریتم FLSPI و $\varepsilon_k \geq 0$ دلخواه باشد به‌طوری‌که $\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|_\infty \leq \varepsilon_k$،

در این صورت

$$Q^{\pi_{k+1}}(s,a) \leq Q^{\pi_k}(s,a) + \frac{2\gamma\varepsilon_k}{1-\gamma}\,, \forall(s,a) \quad (۲۹)$$

اثبات- تعریف می‌کنیم

$$e_k = \sup_{(s,a)}(Q^{\pi_{k+1}}(s,a) - Q^{\pi_k}(s,a))$$

بنابراین

$$Q^{\pi_{k+1}}(s,a) \leq Q^{\pi_k}(s,a) + e_k\,, \forall s,a$$

اما چون $T_\pi Q^\pi = Q^\pi$ و با استفاده از (۲۶):

$$\begin{aligned} Q^{\pi_{k+1}}(s,a) &= T_{\pi_{k+1}}Q^{\pi_{k+1}}(s,a) \\ &\leq T_{\pi_{k+1}}(Q^{\pi_k}(s,a) + e_k) \\ &= T_{\pi_{k+1}}Q^{\pi_k}(s,a) + \gamma e_k \end{aligned}$$

از طرفی در LSPI به دلیل عدم نیاز به نمایش تقریب سیاست، همه خطاهای عملگر صفر هستند [۱]. لذا برای هر k داریم:

$$T_{\pi_{k+1}}Q^{\pi_k} = T_{\pi_k}Q^{\pi_k}.$$

بنابراین:

$$\begin{aligned} Q^{\pi_{k+1}}(s,a) - Q^{\pi_k}(s,a) &\leq T_{\pi_{k+1}}Q^{\pi_k}(s,a) + \gamma e_k \\ &\quad - Q^{\pi_k}(s,a) \\ &= T_{\pi_{k+1}}Q^{\pi_k}(s,a) - T_{\pi_{k+1}}\hat{Q}^{\pi_k}(s,a) \\ &\quad + T_{\pi_{k+1}}\hat{Q}^{\pi_k}(s,a) - Q^{\pi_k}(s,a) \\ &\quad + \gamma e_k \end{aligned}$$

در نتیجه با استفاده از (۲۸) می‌توان نوشت:

$$\begin{aligned} T_{\pi_{k+1}}Q^{\pi_k}(s,a) - T_{\pi_{k+1}}\hat{Q}^{\pi_k}(s,a) + T_{\pi_{k+1}}\hat{Q}^{\pi_k}(s,a) &\\ - Q^{\pi_k}(s,a) + \gamma e_k &\\ \leq \gamma\left|\hat{Q}^{\pi_k}(s,a) - Q^{\pi_k}(s,a)\right| &\\ + \gamma\left|\hat{Q}^{\pi_k}(s,a) - Q^{\pi_k}(s,a)\right| + \gamma e_k &\\ \leq 2\gamma\varepsilon_k + \gamma e_k & \end{aligned}$$

لذا داریم:

$$\sup_{(s,a)}(Q^{\pi_{k+1}}(s,a) - Q^{\pi_k}(s,a)) \leq 2\gamma\varepsilon + \gamma e_k$$

$$\Rightarrow e_k \leq 2\gamma\varepsilon_k + \gamma e_k$$

$$\Rightarrow e_k \leq \frac{2\gamma}{1-\gamma}\varepsilon_k.$$ ■

لم ۵- فرض کنید $\boldsymbol{\pi_k}$ سیاست تولید شده توسط k امین تکرار الگوریتم FLSPI و $\boldsymbol{\varepsilon_k \geq 0}$ دلخواه باشد به‌طوری‌که

$$\|\hat{Q}^{\pi_k} - Q^{\pi_k}\|_\infty \leq \varepsilon_k$$

1 Discount problem

و همچنین فرض کنید

$$f_k = \sup_{(s,a)}(\boldsymbol{Q^{\pi_k}}(s,a) - \boldsymbol{Q^*(s,a)})$$

در این صورت خواهیم داشت:

$$f_{k+1} \le \gamma f_k + \gamma e_k + 2\gamma\varepsilon_k \qquad (۳۰)$$

اثبات- با توجه به فرض داریم

$$Q^{\pi_k}(s,a) \le Q^*(s,a) + f_k, \quad \forall(s,a)$$

اما T عملگری غیر نزولی است [۱۱]، پس با استفاده از (۲۵) و این که $TQ^* = Q^*$:

$$\begin{aligned} TQ^{\pi_k}(s,a) &\le T(Q^*(s,a) + f_k) = TQ^*(s,a) + \gamma f_k \\ &= Q^*(s,a) + \gamma f_k \end{aligned}$$

از طرفی اگر $|a-b| < \varepsilon$، آنگاه $a < b + \varepsilon$ و $b < a + \varepsilon$. با استفاده از این مطلب در فرض قضیه و (۲۵):

$$\begin{aligned} T_{\pi_{k+1}}Q^{\pi_k}(s,a) &\le T_{\pi_{k+1}}\left(\hat{Q}^{\pi_k}(s,a) + \varepsilon_k\right) \\ &= T_{\pi_{k+1}}\hat{Q}^{\pi_k}(s,a) + \gamma\varepsilon_k \\ &= T\hat{Q}^{\pi_k}(s,a) + \gamma\varepsilon_k \\ &\le T(Q^{\pi_k}(s,a) + \varepsilon_k) + \gamma\varepsilon_k \\ &= TQ^{\pi_k}(s,a) + \gamma\varepsilon_k + \gamma\varepsilon_k \\ &\le Q^*(s,a) + \gamma f_k + 2\gamma\varepsilon_k \end{aligned}$$

بنابراین

$$\begin{aligned} Q^{\pi_{k+1}}(s,a) &= T_{\pi_{k+1}}Q^{\pi_{k+1}}(s,a) \\ &\le T_{\pi_{k+1}}(Q^{\pi_k}(s,a) + e_k) \\ &= T_{\pi_{k+1}}Q^{\pi_k}(s,a) + \gamma e_k \\ &\le Q^*(s,a) + \gamma f_k + 2\gamma\varepsilon_k + \gamma e_k \end{aligned}$$

$$\begin{aligned} \Rightarrow\ & Q^{\pi_{k+1}}(s,a) - Q^*(s,a) \\ &\le \gamma f_k + \gamma e_k + 2\gamma\varepsilon_k, \forall(s,a) \end{aligned}$$

$$\Rightarrow f_{k+1} \le \gamma f_k + \gamma e_k + 2\gamma\varepsilon_k. \qquad \blacksquare$$

قضیه ۱- فرض کنید $\pi_1, \dots, \pi_k$ دنباله ای از سیاست‌های تولید شده توسط FLSPI باشد، در این صورت:

$$\boldsymbol{limsup}_{k\to\infty}\left\|\hat{Q}^{\pi_k} - Q^*\right\|_\infty \le \frac{1+\gamma^2}{(1-\gamma)^2}\varepsilon \qquad (۳۱)$$

که در آن $\varepsilon = \boldsymbol{limsup}_{k\to\infty}\varepsilon_k$.

اثبات- با استفاده از لم ۴ داریم

$$\boldsymbol{limsup}_{k\to\infty} f_k \le \gamma\,\boldsymbol{limsup}_{k\to\infty} f_k + \gamma\,\boldsymbol{limsup}_{k\to\infty} e_k + 2\gamma\varepsilon$$

اکنون با استفاده از لم ۳ داریم:

$$\boldsymbol{limsup}_{k\to\infty} e_k \le \frac{2\gamma}{1-\gamma}\varepsilon$$

بنابراین

$$(1-\gamma)\,\boldsymbol{limsup}_{k\to\infty} f_k \le \gamma\frac{2\gamma}{1-\gamma}\varepsilon + 2\gamma\varepsilon = \frac{2\gamma}{1-\gamma}\varepsilon$$

$$\Rightarrow \boldsymbol{limsup}_{k\to\infty} f_k \le \frac{2\gamma}{(1-\gamma)^2}\varepsilon$$

$$\boldsymbol{limsup}_{k\to\infty}\|Q^{\pi_k} - Q^*\|_\infty \le \frac{2\gamma}{(1-\gamma)^2}\varepsilon$$

از طرفی

$$\begin{aligned} \left\|\hat{Q}^{\pi_k} - Q^*\right\|_\infty &= \left\|\hat{Q}^{\pi_k} - Q^{\pi_k} + Q^{\pi_k} - Q^*\right\|_\infty \\ &\le \left\|\hat{Q}^{\pi_k} - Q^{\pi_k}\right\|_\infty \\ &\quad + \|Q^{\pi_k} - Q^*\|_\infty \end{aligned}$$

$$\begin{aligned} \Rightarrow\ & \boldsymbol{limsup}_{k\to\infty}\left\|\hat{Q}^{\pi_k} - Q^*\right\|_\infty \\ &\le \boldsymbol{limsup}_{k\to\infty}\left\|\hat{Q}^{\pi_k} - Q^{\pi_k}\right\|_\infty \\ &\quad + \boldsymbol{limsup}_{k\to\infty}\|Q^{\pi_k} - Q^*\|_\infty \\ &\le \varepsilon + \frac{2\gamma}{(1-\gamma)^2}\varepsilon \end{aligned}$$

$$\Rightarrow \boldsymbol{limsup}_{k\to\infty}\left\|\hat{Q}^{\pi_k} - Q^*\right\|_\infty \le \frac{1+\gamma^2}{(1-\gamma)^2}\varepsilon. \qquad \blacksquare$$

به طور خلاصه دو مهم در این بخش مورد تحلیل قرار گرفته است. آنچه که از لم ۱ حاصل می‌شود این است که مجموعه توابع تعریف شده در سیستم فازی الگوریتم FLSPI، قدرت تقریب هر تابع پیوسته دلخواه را دارند. به بیان روشن‌تر، FLSPI می‌تواند هر تابع ارزش حالت-عمل را با دقت دلخواه، تقریب بزند. در ادامه، آنچه که در لم‌های ۱ تا ۵ آمده است، در جهت آماده کردن مقدمات مربوط به اثبات قضیه ۱ می‌باشد. این قضیه، کران خطایی برای اختلاف تخمین تابع ارزش حالت-عمل، نسبت به تابع ارزش حالت-عمل بهینه، تعریف نموده است. این کران خطا، به حد دنباله ε_k، یعنی دقت تخمین kامین تابع ارزش حالت-عمل تولید شده توسط الگوریتم FLSPI وابسته است. با نزدیک شدن این حد به صفر، کران خطای مزبور نیز به صفر نزدیک شده و تخمین تابع ارزش حالت-عمل تولید شده توسط این الگوریتم، به مقدار بهینه تابع ارزش حالت-عمل نزدیک می‌شود.

۵- شبیه سازی

در این بخش، برای نمایش کارایی روش FLSPI، این روش را در مساله قایق [۲] پیاده سازی کرده‌ایم. هدف، هدایت قایق با شروع از هر موقعیت در سمت چپ رودخانه ای با جریان آب غیر خطی قوی به اسکله سمت راست است. این مسأله دارای دو متغیر حالت پیوسته x و y (مختصات دماغه قایق) در دامنه ۰ تا ۲۰۰ است . مرکز اسکله در موقعیت (۲۰۰،۱۰۰) قرار دارد و عرض آن برابر با ۵ می‌باشد. فضای عمل این مساله، زوایای سکان قایق نسبت به محور افقی است. هدف استفاده از یادگیری تقویتی در این مساله، یاد گرفتن زاویه (عمل) مناسب در هر حالت (موقعیت قایق در رودخانه با جریان آب مربوط به آن موقعیت)، با استفاده از سیگنال تقویتی (پاداش) می‌باشد.

پنج مجموعه فازی برای تقسیم بندی متغیرهای ورودی x و y در نظر گرفته شده است. لذا ۲۵ قاعده فازی خواهیم داشت. خروجی کنترلگر زاویه حرکت قایق می‌باشد. برای تالی هر قاعده دوازده عمل (زاویه حرکت) گسسته لحاظ گردیده است: $A = \{-100,-90,-75,-60,-45,-35,-15,0,15,45,75,90\}$
البته می‌توان برای هر قاعده مجموعه عمل های گسسته متفاوتی در نظر گرفت، لیکن در اینجا یکسان فرض می‌شوند. همچنین مقادیر اولیه وزن (w) برای عمل ها یکسان و برابر صفر لحاظ شده است. کنترل گر با ترکیب این عمل های گسسته، خروجی پیوسته را تولید می کند. هدف از یادگیری پیدا کردن عمل مناسب برای تالی قواعد است.

نتایج هر آزمایش متوسط ارزیابی انجام شده از ۱۰۰ اجرا است. هر اجرا شامل دو بخش یادگیری و آزمون می‌باشد. برای بخش یادگیری، ابتدا ۱۰۰ دسته موقعیت تصادفی تولید می‌شود. هر دسته شامل ۵۰۰۰ نقطه تصادفی شروع حرکت قایق (x=10,y=random) می‌باشد. اگر تعداد اپیزودها (منظور از یک رویداد شروع از نقطه آغاز و رسیدن به طرف دیگر ساحل است) از ۵۰۰۰ بیشتر شود یا تعداد دفعات متوالی که قایق به ناحیه غیر شکست رسیده است به ۴۰ برسد بخش یادگیری پایان می یابد. بخش آزمون شامل ۴۰ رویداد با شروع از ۴۰ نقطه با x=10 و y پخش شده با فواصل مساوی در رنج ۰ تا ۲۰۰ است.

نتایج حاصل از ۱۰۰ اجرای مجزای این الگوریتم در جدول ۱ با بهترین نتیجه‌های روش‌های FQL و FSL [۸] مقایسه شده است. در این جدول، متوسط تعداد اپیزود در یادگیری(Avg. LDI) ، انحراف معیار اپیزودها در یادگیری(std. LDI) ، میانگین فاصله در آزمایش (Avg. DEI) ، نرخ شکست[1] در آزمایش و همچنین میانگین زمان اجرا (Avg. time) آمده است. سیستم کامپیوتری مورد استفاده در این آزمایش دارای پردازشگر intel core i7 (2.20GH) و ۸ گیگابایت حافظه می‌باشد.

همانطور که در جدول ۱ مشاهده می‌شود، FLSPI نتیجه بسیار بهتری نسبت به FQL و FSL دارد. برای مثال، FLSPI به ترتیب نسبت به FQL و FSL، از نظر زمان آموزش (LDI)، ۲ و ۲/۸ برابر، از نظر کیفیت یادگیری (DEI)، ۲/۸ و ۱/۵ برابر و از لحاظ نرخ شکست، ۵/۷ و ۲/۴ برابر بهتر شده است. از لحاظ زمان اجرا نیز مشاهده می‌شود که FLSPI نسبت به FQL و FSL به ترتیب ۵/۸ و ۲/۹ برابر سریع‌تر است.

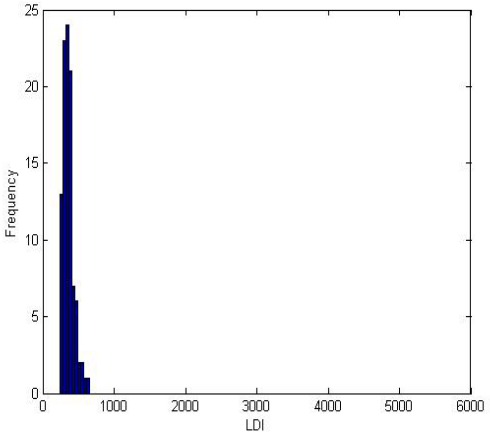
جدول ۱ : مقایسه نتایج شبیه سازی

Initial parameters	Method	Avg. DEI	Avg. LDI	Std. (LDI)	Failure Rate	Avg. Time (Sec)
$\delta = 0.01$ $\alpha = 0.01$	FQL	8.69	733	1084	6.7	83.81
$\delta = 1$ $\alpha = 0.1$	FSL	4.45	1010	1065	2.825	41.97
	FLSPI	3.06	360.9	76.04	1.17	14.45

هیستوگرام یادگیری FLSPI و FSL در شکل‌های ۱ و ۲ آمده است. همانطور که دیده می‌شود در FLSPI، به ازای ۱۰۰ اجرای متمایز، هیچ مورد واگرایی وجود ندارد. و در همه اجراها، یادگیری با تعداد اپیزود کمتر از ۶۷۰ صورت گرفته است.

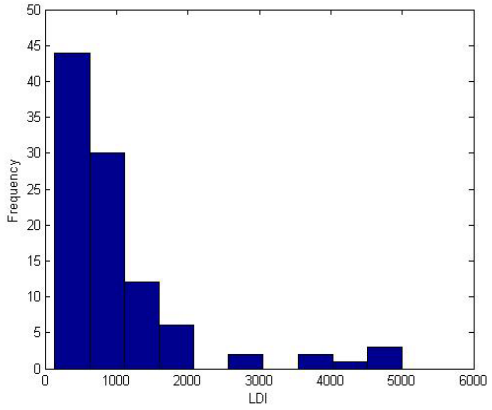
در شکل ۳ تغییرات وزن‌های قاعده اول در اپیزودهای متوالی، از این پیاده سازی برای FLSPI نشان داده شده است. همانطور که دیده می‌شود تغییرات وزن‌ها بعد از حدود ۳۰۰ تکرار بسیار کم شده است و می توان گفت که بعد از حدود ۱۰۰۰ تکرار تقریبا بدون تغییر هستند. لذا وزن‌ها در FLSPI خیلی سریع همگرا می‌شوند.

برای نمونه، نتیجه آزمایش الگوریتم FLSPI روی ۴۰ داده آزمایشی در شکل ۴ آمده است. همانطور که ملاحظه می‌شود یادگیری به صورت مطلوبی اتفاق افتاده است. در واقع یادگیری به ازای همه نقاط فضا به صورت کامل می‌باشد.

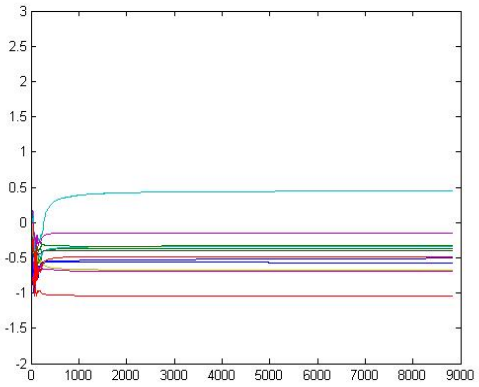


شکل ۱: هیستوگرام یادگیری برای FLSPI

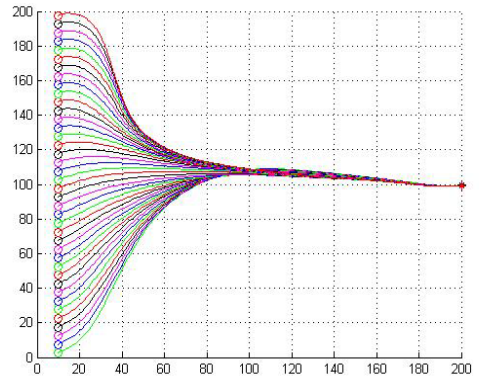
1 Failure Rate



شکل ۲: هیستوگرام یادگیری برای FSL



شکل ۳: تغییرات وزن‌های قاعده اول FLSPI



شکل ۴: نتیجه الگوریتم FLSPI پس از آموزش، روی داده‌های تست

۵- ٦-نتیجه‌گیری

در این مقاله یک روش جدید یادگیری تقویتی فازی با استفاده از LSPI ارائه شد. الگوریتم ارائه شده FLSPI نامیده شد. نحوه تعریف توابع پایه ارزش- عمل، و بروز رسانی پارامترهای وزن در FLSPI شرح داده شد.

ابتدا اثبات شد که FLSPI شرایط مربوط به قضایای بیان شده در LSPI را بر آورده نموده و لذا نتایج آن قضایا برای FSLPI صادق است. همچنین اثبات شد که خطای تابع ارزش عمل تقریب زده شده با مقدار واقعی محدود بوده و کمتر از مقدار کرانی است که در این مقاله بدست آمده است. جهت ارزیابی FLSPI، عملکرد آن در مساله شناخته شده قایق با دو روش یادگیری تقویتی فازی نقاد-تنها به نام های FQL و FSL مقایسه شد. از دیگر مزیت‌های روش ارائه شده، عدم وابستگی آن به نرخ یادگیری می‌باشد. همچنین عدم وجود واگرایی، از ویژگی‌های قابل توجه FLSPI می‌باشد.نتایج شبیه سازی نشان داد که همگرایی وزن‌ها در این الگوریتم خیلی سریع اتفاق می‌افتد. در ضمن این‌که زمان اجرای این الگوریتم، آن را برای مسائلی که نیاز به الگوریتم های بلادرنگ دارند، مناسب می‌کند.

مراجع

[1] Lagoudakis M. G., and Parr, R., "Least-squares policy iteration", Journal of Machine Learning Research, vol. 4, p. 1107-124, 2003.

[2] Jouffe, L., "Fuzzy inference system learning by reinforcement methods", IEEE Trans. Syst., Man, Cybern. C, vol. 28, p. 338–35, 1998.

[3] Berenji, H., " Fuzzy Q-learning: a new approach for fuzzy dynamic programming", IEEE World Congress on Computational Intelligence, 1994.

[4] Panahi, F.H., Ohtsaki, T., "Optimal Channel-Sensing Scheme for Cognitive Radio Systems Based on Fuzzy Q-Learning", IEICE TRANSACTIONS on Communications, Vol.E97-B, No.2, pp.283-29, 2014.

[5] Shamshirband, S., Patelc, A., Anuarb, N.B., Mat Kiahb, M.L., Abrahame, A., "Cooperative game theoretic approach using fuzzy Q-learning for detecting and preventing intrusions in wireless sensor networks", Engineering Applications of Artificial Intelligence, 2014.

[6] Muñoz, P., Barco, R., De la Bandera, I., "Optimization of load balancing using fuzzy Q-Learning for next generation wireless networks", Expert Systems with Applications Volume 40, Issue 4, Pages 984–994, March 2013.

[7] Glorennec, P.Y. and Jouffe, L., "Fuzzy Q-learning", Proc. IEEE Int. Conf. Fuzzy Systems, vol 2, p. 659-662, July 1997.

[8] Derhami, V., Majd , V.J., and Ahmadabadi, M.N., "Fuzzy sarsa learning and the proof of existence of its stationary points", Asian Journal of Control, vol 10, p. 535-549, September 2008.

[9] Rovcanin M., De Pooreter, E., Moerman, I., Demeester, P., "A reinforcement learning based solution for cognitive network cooperation between co-located, heterogeneous wireless sensor networks", Ad Hoc Netw, vol 17, p. 98-113, June 2014.

[10] Thiery, C., Scherrer, B.L., "east-Squares λ Policy Iteration: Bias-Variance Trade-off in Control Problems", International Conference on Machine Learning, 2010.

[11] Bertsekas, D.P., and Tsitsiklis, J.N., "Neuro-Dynamic Programming", Athena Scientific, Belmont, Massachusetts, 1996.

[12] Bus¸oniu, L., Ernst, D., De Schutter, B., "Online least-squares policy iteration for reinforcement learning control", American Control Conference (ACC-10), Baltimore, US, 30 June – 2 July 2010.

[13] Jang, J.S.R., Sun, C.T., and Mizutani, E., "Neuro-Fuzzy and soft computing", Prentice-Hall, 1997.

[14] Razavi, R., Klein, S., and Claussen, H., "Self-Optimization of Capacity and Coverage in LTE Networks Using a Fuzzy Reinforcement Learning Approach", 21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2010.

[15] Castro, J.L., "Fuzzy logic controllers are universal approximators", vol 25, p. 629-63, April 1995.

[16] Du, K.L., Swamy, M. N. S., "Reinforcement Learning. Neural Networks and Statistical Learning", pp 547-561, 2014.

[17] Howard, R.A., "Dynamic Programming and Markov Processes", MIT Press, Cambridge, Massachusett, 1960.