اولین همایش ملی علوم و مهندسی کامپیوتر و فناوری اطلاعات
1st National Conference on Computer Science, Engineering and Information Technology
دانشگاه آزاد اسلامی واحد شیراز
۱۱ آبان ماه ۹۱

یادگیری سارسا فازی با توزیع محلی پاداش

سلمان سلطانیان [۱] –محمد باقر نقیبی سیستانی [۲]
(۱) گروه برق – دانشگاه فردوسی مشهد
sa.soltanian@stu-mail.um.ac.ir
(۲) گروه برق – دانشگاه فردوسی مشهد
mb_naghibi@um.ac.ir

چکیده

یادگیری در محیط های پیوسته به علت اهمیت آن در محیط های واقعی یکی از مسائل مهم در یادگیری تقویتی می باشد و ارائه روش های یادگیری سریعتر، مقاومتر و در عین حال با همگرایی بهتر همچنان یکی از دغدغه های اصلی پژوهشگران در این حوزه است. در این راستا در این مقاله یک روش یادگیری جدید با نام یادگیری سارسا فازی با توزیع محلی پاداش (FSL-LDR) ارائه شده است. این روش توانایی یادگیری در محیط های پیوسته را داشته و عمل پیوسته تولید می کند. عملکرد این روش در مسئله قایق از دو منظر سرعت یادگیری و سیاست نهایی مورد ارزیابی قرار گرفته و با روش های یادگیری- Q فازی و یادگیری سارسا فازی مقایسه شده است. نتایج نشان می دهند که این روش عملکرد بهتری نسبت به روشهای دیگر دارد. سپس در مسئله تعادل آونگ- ارابه امکان واگرایی روش یادگیری-Q فازی نشان داده شده است.

کلمات کلیدی

یادگیری ماشین، یادگیری تقویتی، کنترل فازی، یادگیری سارسا، یادگیری-Q، توزیع پاداش.

۱ - مقدمه

یادگیری تقویتی روشی جدید و قدرتمند برای یافتن سیاست کنترل بهینه در محیط های نامشخص می باشد. بسیاری از روش های سنتی یادگیری تقویتی تنها برای مسائل با محیط گسسته طراحی شده اند. یادگیری در این مسائل به علت نویز، پاداش تاخیری، محیط نامعلوم و تصادفی مشکل می باشد. با این وجود در بسیاری از مسائل واقعی فضای حالت و عمل بسیار وسیع یا پیوسته بوده و به علت پدیده نفرین ابعاد، یادگیری مشکلتر و پیچیده تر می شود. به منظور حل این مشکل از تقریب تابع[1] در یادگیری تقویتی استفاده می گردد تا بتوان تجربیات را به حالات مجاور تعمیم داد[۱].

سیستمهای فازی تقریب زن های جهانی می باشند. این سیستم ها دارای توانایی زیادی از جمله نمایش دانش توسط قوانین اگر-آنگاه، مدل سازی و کنترل سیستم های غیر خطی و نامعین با دقت دلخواه می باشند[۲]. تلفیق سیتسم فازی و یادگیری تقویتی باعث بوجود آمدن دسته جدیدی از روشها با نام یادگیری تقویتی فازی (FRL) شده است. [۳] با تلفیق کردن روش یادگیری-Q و سیستم فازی، روش یادگیری- Q فازی (FQL-I) را ارائه داد. این روش بر روی چند مسئله نیز پیاده سازی شده است [۳, ۴]. [۵] امکان واگرایی این روش را با یک مثال نشان داد و روش یادگیری سارسا فازی (FSL) را بر پایه آن ارائه نمود و همگرایی آن را تحت سیاست ایستا و برای محیط گسسته اثبات کرد.

بنارینی و همکارانش[۶] گامی مهم در تحلیل روش FQL نهادند. آنها مزایا و معایب این روش را بررسی نموده و روشی جدید (FQL-II) برای محاسبه ارزش عمل نهایی ارائه کردند. در این روش توزیع پاداش به صورت محلی صورت می گیرد و این امر موجب می شود که این روش شباهت بیشتری به روش یادگیری- Q سنتی پیدا کند و بروز رسانی ارزش همگن تر شده، واریانس کمتری داشته باشد و در نتیجه میزان نوسانات ارزش کمتر و فرایند یادگیری کاراتر گردد.

در این مقاله روش یادگیری سارسا فازی با توزیع محلی پاداش (FSL-LDR) را ارائه می دهیم. به منظور استفاده بهتر از اطلاعات، از شایستگی پیگردی[2] در این روش استفاده می نماییم. عملکرد این روش را در مثال قایق مورد ارزیابی قرار داده و با روشهای دیگر مقایسه می کنیم. با طرح یک آزمایش نشان خواهیم داد که روش [۶]، علی رغم

[1] Function approximation

[2] Eigibility traces

استفاده از یادگیری-Q (که یک روش برون سیاست[1] است) برون-سیاست نمی باشد و در نهایت در مثال تعادل آونگ- ارابه امکان واگرایی این روش را نشان می دهیم.
در بخش دوم، یادگیری تقویتی به طور مختصر معرفی می شود. بخش سوم به تشریح روش پیشنهادی پرداخته و در بخش چهارم نتایج شبیه سازی و مقایسه آورده شده است. مقاله با نتیجه گیری و بیان کارهای آینده به پایان می رسد.

۲ – یادگیری تقویتی

ساختار متداول یادگیری تقویتی از دو بخش عامل و محیط تشکیل شده است[۷]. عامل در هر گام زمانی t، حالت محیط x^t را دریافت می نماید و عمل a^t را به محیط اعمال می کند. این عمل، محیط را به حالت جدید x^{t+1} برده و حالت جدید به همراه سیگنال پاداش r^{t+1} به عامل داده می شود. هدف در یادگیری تقویتی یافتن نگاشتی از حالت به عمل به منظور بیشینه کردن پاداش در دراز مدت است. به این نگاشت، سیاست بهینه گفته می شود. متناظر با هر سیاست π یک تابع ارزش حالت و عمل $Q^\pi(x,a)$ وجود دارد که نشان دهنده امید ریاضی مجموع تنزیلی پاداشهای دریافتی در دراز مدت با شروع از حالت x و انجام عمل a و دنبال کردن سیاست π می باشد.

$$Q^\pi(x,a)=E\left\{\sum_{k=0}^{\infty}\gamma^k r^{t+k+1}\middle|\pi,x^t=x,a^t=a\right\} \quad (۱)$$

γ ضریب تضعیف[2] نامیده می شود. سیاست بهینه، مقدار تابع Q را بیشینه می کند. لذا بسیاری از روش های یادگیری تقویتی به جای محاسبه مستقیم سیاست بهینه، تابع ارزش بهینه Q^* را یافته و سیاست بهینه را از روی آن بدست می آورند.

۳ – یادگیری سارسا فازی با توزیع محلی پاداش

در این بخش، روش پیشنهادی به طور کامل معرفی می شود. در این روش از یک سیستم فازی تاکاگی سوگینو نوع صفر برای نگاشت حالت $x=(x_1,\ldots,x_N)\in X=X_1\times\ldots\times X_N$ به عمل استفاده شده است. در هر بعد k فضای حالت n_{X_k} مجموعه فازی تعریف شده است. سیستم فازی تاکاگی سوگینو با n_R قانون به فرم زیر را در نظر بگیرید.

$$\begin{aligned}&R_r:\text{If } x_1 \text{ is } L_{j1} \text{ and } \ldots \text{ and } x_N \text{ is } L_{jN} \text{ Then}\\ &y=a_{r,1} \text{ with } q_{r,1} \text{ or } \ldots \text{ or } y=a_{r,n_r} \text{ with } q_{r,n_r}\end{aligned} \quad (۲)$$

که L_{jk} مجموعه فازی jk ام بر روی X_k می باشد و $jk=1,\ldots,n_{X_k}$.

به هر قانون R_r یک مجموعه n_r تایی عمل گسسته $A_r=\{a_{r,1},a_{r,2},\ldots,a_{r,n_r}\}$ و یک مجموعه ارزش عمل متناظر با آن $Q_r=\{q_{r,1},q_{r,2},\ldots,q_{r,n_r}\}$ نسبت داده شده است.
T-norm به کار رفته برای اجرای اپراتور and، عمل ضرب می باشد.
هر قانون ناحیه ای از فضای حالت را پوشش می دهد که حالت فازی می نامیم. ورودی x^t می تواند به یک یا چند حالت فازی تعلق داشته باشد. به این قوانین، قوانین فعال گفته می شود. درجه فعالیت قانون R_r به صورت زیر محاسبه می شود.

$$\phi_r(x^t)=\prod_{i=1}^{N}\mu_{L_{ji}}(x_i) \quad (۳)$$

$\mu_{L_{ji}}(x_i)$ میزان تعلق x_i به مجموعه فازی L_{jk} است.

الف- انتخاب عمل

به منظور تعیین عمل نهایی در حالت x^t، رقابتی بین عمل های گسسته درهر قانون فعال صورت می گیرد و عمل برنده در هر قانون فعال تعیین می شود. اپراتور or در (۲) نشان دهنده این رقابت است. عمل نهایی $A^t(x^t)$ از ترکیب خطی عمل های برنده قوانین فعال به نسبت درجه فعالیت نرمالیزه شده آنها بدست می آید:

$$A^t(x^t)=\sum_{r=1}^{n_R}\psi_r(x^t)\breve{a}_r^t \quad (۴)$$

$\breve{a}_r^t$، عمل برنده در قانون r ام می باشد که با استراتژی کاوش اپسیلون گریدی[3] از بین عمل های گسسته این قانون و با توجه به ارزش متناظر با این عمل ها انتخاب می شود. درجه فعالیت نرمالیزه شده قانون r ام به صورت زیر تعریف می شود:

$$\psi_r(x^r)=\frac{\phi_r(x^t)}{\sum_{r=1}^{n_R}\phi_r(x^t)} \quad (۵)$$

عمل نهایی می تواند مقادیر پیوسته داشته باشد. ارزش عمل نهایی $Q^t(x^t,A^t(x^t))$ از درون یابی خطی ارزش عمل نهایی در قوانین فعال محاسبه می شود:

---

[1] Off-policy
[2] Discount factor
[3] $\varepsilon-greedy$

$$Q^t(x^t, A^t(x^t)) = \sum_{r=1}^{n_R} \psi_r(x^t)\tilde{Q}_r(A^t(x^t)) \qquad (۶)$$

$\tilde{Q}_r(A^t(x^t))$ مقدار ارزش عمل نهایی در قانون r ام می باشد که از درون یابی خطی ارزش عمل های مجاورِ عمل نهایی در قانون rام بدست می اید. اگر فرض کنیم $a^t_{r,i_r}, a^t_{r,i_r+1}$ عملهای مجاور عمل نهایی در قانون rام در گام زمانی t باشند، داریم:

$$\tilde{Q}_r(A^t(x^t)) = \mu^t_{r,i_r} q^t_{r,i_r} + \mu^t_{r,i_r+1} q^t_{r,i_r+1} \qquad (۷)$$

$$\mu^t_{r,i_r} = \left|\frac{A^t(x^t) - a^t_{r,i_r+1}}{a^t_{r,i_r+1} - a^t_{r,i_r}}\right|,\ \mu^t_{r,i_r+1} = \left|\frac{A^t(x^t) - a^t_{r,i_r}}{a^t_{r,i_r+1} - a^t_{r,i_r}}\right| \qquad (۸)$$

[۳, ۵] ارزش عمل نهایی را از درون یابی ارزش عمل های برنده در قوانین فعال بدست می آورند. در صورتیکه در این روش ارزش عمل نهایی از درون یابی ارزش عمل های محلی در قوانین فعال بدست می آید. این روش محاسبه موجب می شود تمام پارامتر ها به خودی خود معنا دار بوده و نماینده محیط اطراف خود باشند، به علاوه اطلاعات بروز رسانی همگن تر و واریانس کمتری داشته باشند و در نتیجه فرایند یادگیری کاراتر گردد [۶].

ب) یادگیری

در مرحله یادگیری بایستی با توجه به پاداش دریافتی و ارزش عمل نهایی بعدی، مقدار ارزش عمل نهایی فعلی $Q(x^t, A^t(x^t))$ بروز شود. در روش های تابعی، مانند یادگیری تقویتی فازی افزایش یا کاهش ارزش معمولا با روش تندترین نشیب[1] انجام می گیرد[۱, ۶].

لازم به ذکر است که عملهای گذشته باعث رسیدن به وضعیت فعلی شده اند و بروز رسانی ارزش عمل های گذشته باعث استفاده بهتر از اطلاعات می گردد. بدین منظور از شایستگی پیگردی[2] [۱] در فرایند یادگیری استفاده می نماییم. به منظور پیاده سازی این روش به هر عمل در تمام قوانین یک مقدار شایستگی نسبت می دهیم. فرض کنید $e^t_{r,j}$ مقدار شایستگی عمل j ام در قانون rام باشد. رابطه بروز رسانی برای تمام قوانین و عمل ها به صورت زیر است:

$$q^{t+1}_{r,j} = q^t_{r,j} + \alpha\delta^t_Q e^t_{r,j} \qquad (۹)$$

$$\delta^t_Q = r^t_{x^t, A^t(x^t)} + \gamma * Q(x^{t+1}, A^{t+1}(x^{t+1})) - Q^t(x^t, A^t(x^t)) \qquad (۱۰)$$

[1] Gradient descent
[2] Eligibility traces

$r^t_{x^t, A^t(x^t)}$ پاداش دریافت شده از محیط پس از اعمال عمل $A^t(x^t)$ می باشد و α نرخ یادگیری[3] است. اگر $a^t_{r,i_r}, a^t_{r,i_r+1}$ عمل های مجاور عمل نهایی در قانون rام باشند، مقادیر شایستگی در تمام قوانین به صورت زیر بروز می شوند:

$$e^{t+1}_{r,j} = \begin{cases} \gamma\lambda e^t_{r,j} + \psi_r(x^t)\mu^t_{r,j} & \text{if } j = i_r \text{ or } j = i_r + 1 \\ \gamma\lambda e^t_{r,j} & \text{otherwise} \end{cases} \qquad (۱۱)$$

λ ضریب فراموشی[4] می باشد.

ج) مراحل اجرا

برای اجرای این روش، در هر گام زمانی، ۸ مرحله زیر بایستی اجرا شود:

۱) دریافت حالت x^{t+1} و پاداش $r^t_{x^t, A^t(x^t)}$ از محیط.
۲) انتخاب عمل در تمام قوانین فعال توسط روش اپسیلون گریدی.
۳) محاسبه عمل نهایی و ارزش عمل نهایی توسط روابط (۴) و (۶).
۴) محاسبه δ^t_Q و بروز رسانی تمام q ها توسط (۱۰) و (۹).
۵) محاسبه مجدد ارزش عمل نهایی توسط (۶).
۶) بروز رسانی ضرایب شایستگی (۱۱).
۷) اعمال عمل نهایی به محیط.
۸) $t \leftarrow t+1$ و بازگشت به مرحله ۱.

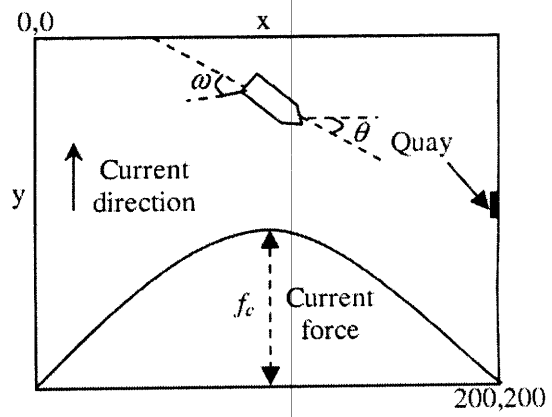
۴ – شبیه سازی

در این بخش ابتدا عملکرد روش FSL-LDR در مسئله قایق مورد ارزیابی قرار گرفته و با روش های FQL-I، FQL-II و FSL مقایسه می شود. سپس امکان واگرایی روش FQL-II در مسئله تعادل آونگ-ارابه نشان داده خواهد شد.

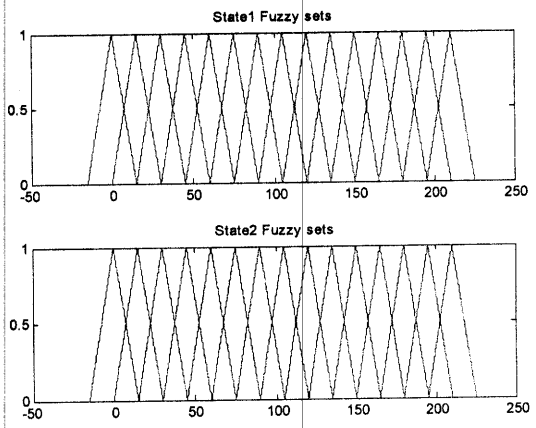
الف) مسئله قایق

در این مسئله از روش های یادگیری تقویتی فازی برای هدایت قایق از ساحل چپ رودخانه به سمت اسکله ای در ساحل راست رودخانه در حضور جریان شدید و غیر خطی آب استفاده می کنیم. هدف رسیدن به اسکله از هر نقطه ای در ساحل چپ رودخانه می باشد (شکل ۱).

[3] Learning rate
[4] Recency factor



شکل۱: مسئله قایق [۵].



شکل ۲: توابع عضویت ورودی

این مسئله دارای دو متغییر حالت پیوسته x و y می باشد. اسکله در مختصات (۱۰۰،۲۰۰) قرار دارد و عرض آن ۵ می باشد. معادلات دینامیک قایق مانند [۵] در نظر گرفته شده است و به دلیل کمبود فضا خوانندگان را به این مقاله ارجاع می دهیم. تابع پاداش بر پایه سه ناحیه ساحلی می باشد: ناحیه موفقیت Z_s که اسکله را شامل می شود $(x=200, y\in[97.5,102.5])$، ناحیه قابل قبول Z_v که $x=200, y\in[92.5,97.5]\cup[102.5,107.5]$ و ناحیه شکست Z_f که بقیه نقاط ساحلی را شامل می شود. تابع پاداش به صورت زیر تعریف شده است[۵]:

$$R(x,y)=\begin{cases}+1 & (x,y)\in Z_s\\ D(x,y) & (x,y)\in Z_v\\ -1 & (x,y)\in Z_f\\ 0 & otherwise\end{cases} \qquad (۱۲)$$

D(x,y) تابعی است که با فاصله گرفتن از ناحیه موفقیت به طور خطی از ۱+ به ۱- تغییر می کند.

۱۵مجموعه فازی مثلثی هر بعد را پوشش داده که در مجموع ۲۲۵

جدول۱: نتایج شبیه سازی برای ضرایب فراموشی مختلف

Recency factor	Method	Avg.Dis	Avg.LSpeed	Std.LSpeed
$\lambda=0$	FQL-II	9.02	1432	1332
	FSL-LDR	8.99	933	1243
$\lambda=0.5$	FQL-II	8.38	643	1123
	FSL-LDR	8.31	538	834
$\lambda=0.9$	FQL-II	7.56	521	1110
	FSL-LDR	5.22	342	678
$\lambda=0$ $\varepsilon=1$	FQL-II	64.3	5000	0

قانون را تشکیل می دهند (شکل ۲). مجموعه عمل های گسسته تمام قوانین یکسان و از ۱۲ جهت مختلف از جنوب به شمال تشکیل شده است[۵].

$A_r=\{-100,-90,-75,-60,-45,-35,-15,0,15,45,75,90\}$

عامل با ترکیب این عمل ها، عمل پیوسته تولید می کند.

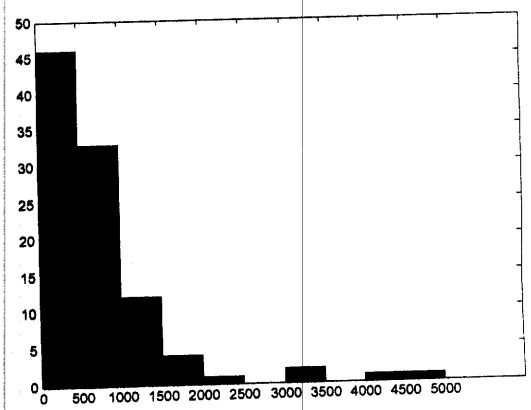
نتیجه هر آزمایش از میانگین گیری معیارهای عملکرد معرفی شده در [۵] در ۱۰۰ اجرا بدست می آید. هر اجرا از یک مرحله یادگیری و یک مرحله تست تشکیل شده است. مرحله یادگیری وقتی پایان می یابد که قایق ۴۰ دفعه پشت سرهم به ناحیه موفقیت یا ناحیه قابل قبول برسد و یا تعداد تلاش [1] ها از ۵۰۰۰ بیشتر گردد (تلاش به گذار از حالت اولیه به ساحل اتلاق می گردد). پس از پایان مرحله یادگیری مرحله تست شروع می شود. این مرحله از ۴۰ تلاش تشکیل شده است. نقاط شروع در هر دو مرحله تصادفی می باشد (x=200 و y تصادفی با توزیع یکنواخت در بازه [0,200]). در مرحله یادگیری مدت یادگیری اندازه گیری می شود (تعداد تلاش ها در پایان مرحله یادگیری) و در مرحله تست میانگین فاصله تا اسکله در۴۰ تلاش محاسبه می شود. فاصله تا اسکله در پایان هر تلاش و از رابطه زیر بدست می آید.

$$d(x,y)=\begin{cases}|y-100| & \text{if right bank reached}\\ 100+(200-x) & \text{otherwise}\end{cases} \qquad (۱۳)$$

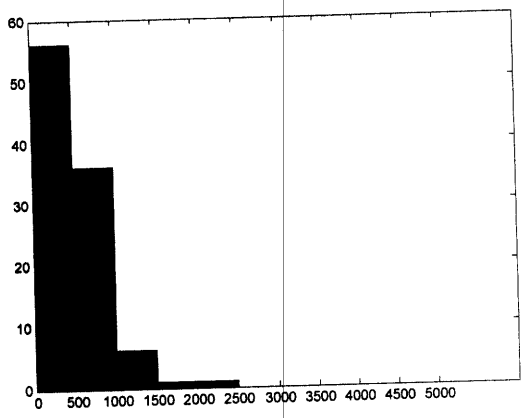
به منظور بررسی تاثیر ضرایب شایستگی، عملکرد روش FQL-II و FSL-LDR برای مقادیر مختلف λ مورد ارزیابی قرار گرفته است.

جدول۱ میانگین مدت یادگیری (Avg.LSpeed)، میانگین فاصله (Avg.Dis) و انحراف معیار مدت یادگیری (Std.LSpeed) را برای این دو روش نشان می دهد. بدین منظور از شایستگی پیگردی در روش FQL-II استفاده نمودیم. نتایج متناظر با $\lambda=0$، مربوط به روش اصلی FQL-II [6] می باشد. نتایج نشان می دهند که شایستگی پیگردی تاثیر بسیاری در کاهش مدت یادگیری و کاهش میزان فاصله در هر دو روش دارد. بهترین نتایج برای $\lambda=0.9$ بدست آمده است. با افزایش بیشتر λ مدت یادگیری افزایش می یابد. در تمام این حالت

[1] Epizode



شکل۳: هیستوگرام مدت یادگیری برای روش FQL-II



شکل۴: هیستوگرام مدت یادگیری برای روش FSL-LDR

ها، روش FSL_LDR بهتر از روش FQL-II عمل کرده است. شکل ۳و ۴ هیستوگرام مدت یادگیری را برای هر دو روش نشان می دهد. همانطور که مشاهده می شود احتمال طولانی شدن یادگیری در روش FSL_LDR کمتر است و در نتیجه این روش، روش قابل اعتمادتری می باشد.

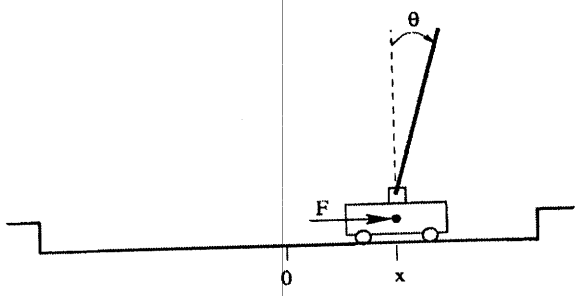
در روش های برون سیاست[1](مانند یادگیری-Q) سیاستی که با محیط تعامل می شود (سیاست تعاملی[2]) مستقل از سیاستی است که ارزش آن تخمین زده می شود[3]. لذا این روش ها می توانند بدون تعامل برخط[4] با محیط و تنها از روی داده های برون خط[5] سیاست بهینه را بیابند و این امر امتیاز مهمی برای این روشها در برخی کاربردها محسوب می شود[۱, ۸]. روش FQL-II از تلفیق روش یادگیری Q با سیستم فازی بوجود آمده است ولی مشخص نیست که آیا این روش برون سیاست است یا خیر.

---

[1] Off-policy
[2] Behavior policy
[3] Estimation policy
[4] Online
[5] Offline

جدول ۲: مقایسه روش های یادگیری در مسئله قایق

Method	Avg.Dis	Avg.LSpeed	Std.LSpeed
FQL-I	8.69	733	1084
FSL	8.68	698	1013
FQL-II	7.56	521	1110
FSL-LDR	5.22	342	678



شکل۵: سیستم آونگ- ارابه[۳].

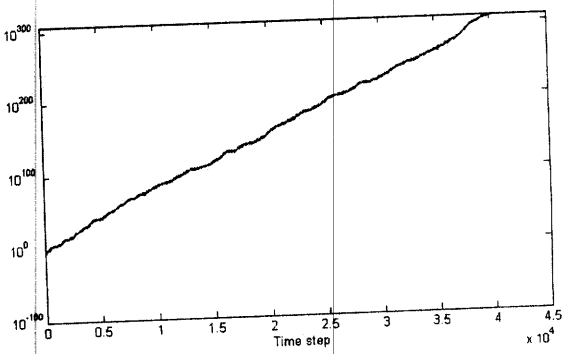
به منظور بررسی این مسئله، سیاست را در این روش ثابت و کاملا تصادفی ($\varepsilon = 1$) قرار دادیم. ردیف ۷ جدول ۱ نتایج این آزمایش را نشان می دهد، همانطور که انتظار می رفت، به علت سیاست کاملا تصادفی یادگیری در تمام اجرا ها تا مرز بالایی(۵۰۰۰ تلاش) ادامه پیدا کرد ولی میانگین فاصله (Avg.Dis) در مرحله تست بسیار زیاد شد. این امر نشان می دهد که این روش نتوانسته است به نزدیک سیاست بهینه برسد. لذا این روش با وجود شباهتی که به روش یادگیری- Q دارد ولی یک روش برون سیاست نمی باشد و از این لحاظ مزیتی نسبت به روش FSL-LDR ندارد.

عملکرد چهار روش FQL-I، FQL-II، FSL و FSL-LDR در جدول ۲ بایکدیگر مقایسه شده است. نتایج مربوط به روش FSL و FQL-I بهترین جواب های این دو روش هستند که در [۵] گزارش شده است. نتایج نشان می دهندکه اختلاف زیادی بین روش های FQL-II، FSL-LDR با روشهای FSL و FQL-I از نظر هر دو معیار عملکرد وجود دارد. علت این اختلاف بروز رسانی همگن تر و با واریانس کمتر است که ناشی از توزیع محلی پاداش در این روشها می باشد.

ب) مسئله تعادل آونگ- ارابه

در این مسئله یک آونگ به صورت وارونه بر روی یک ارابه قرار گرفته است. ارابه می تواند آزادانه در یک بعد حرکت نماید. کنترل کننده نیروی کنترلی را در بازه های زمانی گسسته به ارابه اعمال می کند (شکل ۵). دینامیک این مسئله مانند [۳] در نظر گرفته شده است.

مرحله یادگیری هنگامی پایان می یابد که ۵۰۰۰ تلاش به پایان برسد و یا طول یک تلاش از ۵۰۰۰۰ گام زمانی بیشتر شود. هر تلاش از یک حالت تصادفی شروع شده و هنگامی پایان می یابد که خطایی رخ دهد و یا تعداد گام های زمانی بیش از حد شود. خطا هنگامی رخ می دهد که زاویه آونگ $|\theta| > 12^\circ$ گردد یا مکان ارابه $|x| > 2.4m$ شود. در این حالت جریمه ۱- به عامل داده می شود.



شکل ۶: مقدار یکی از پارامتر های سیستم فازی در حین اخرین تلاش در روش FQL-II

عمل های گسسته تمام قوانین یکسان و برابر $A_r = \{-10,-8,-6,-4,-2,0,2,4,6,8,10\}$ در نظر گرفته شد. تعداد قوانین فازی ۳۲۰ عدد (پنج مجموعه فازی برای زاویه، چهار مجموعه فازی برای سرعت زاویه ای، موقعیت و سرعت خطی) توابع فعالیت تمام مجموعه های فازی به صورت مثلثی می باشد و این توابع به صورتی هستند که در هر بعد، فازی سازی قوی بوده و قاعده توابع عضویت در هر بعد یک اندازه می باشد. پارامتر های یادگیری $\gamma = 0.9, \alpha = 0.4, \varepsilon = 0.4$ در نظر گرفته شد. مشاهده شد که پارامتر های سیستم فازی(ارزش عمل های گسسته قوانین) واگرا شدند. شکل ۶ واگرایی یکی از این پارامتر ها را نشان می دهد.

لازم به ذکر است که عملکرد روش FSL-LDR نیز در هر دو مثال برای مقادیر مختلف پارامتر ها مورد بررسی قرار گرفت ولی این روش در تمام موارد همگرا بود.

۵ – نتیجه گیری

یک روش جدید در حوزه یادگیری تقویتی با نام یادگیری سارسا فازی با توزیع محلی پاداش (FSL-LDR) ارائه دادیم. این روش قابلیت یادگیری در مسائل با حالت و عمل پیوسته را داراست. عملکرد این روش در مسئله قایق با روش های دیگر مقایسه شد. نتایج شبیه سازی نشان دادند استفاده از یادگیری سارسا به جای یادگیری Q علاوه بر افزایش سرعت یادگیری، باعث همگرایی بهتر این روش گردیده است. همچنین استفاده از شایستگی پیگردی تاثیر زیادی در افزایش سرعت همگرایی این روش دارد. امکان واگرایی روش یاگیری- Q فازی در مسئله تعادل آونگ- ارابه نشان داده شد از سوی دیگر شبیه سازی به ازای مقادیر مختلف پارامتر ها هیچ مورد واگرایی روش FSL-LDR را نشان ندادند. سیستم فازی استفاده شده در این روش یک سیستم خطی می باشد و با توجه به مزیت روشهای خطی در تحلیل ریاضی همگرایی، این امکان وجود دارد که بتوان همگرایی این روش را اثبات نمود.

در این روش به علت توزیع محلی پاداش ها، هر یک از پارامتر های سیستم فازی به خودی خود معنادار می باشند، در نتیجه اضافه کردن یا حذف یک قانون اختلالی در یادگیری های قبلی ایجاد نمی کند. لذا این روش گزینه مناسبتری برای استفاده در روشهای تطبیقی می باشد.

مراجع

[1] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*: The MIT Press, 1998.

[2] L. Wang, *A course in fuzzy systems and control*: Prentice-Hall, Inc., 1997.

[3] L. Jouffe, "Fuzzy inference system learning by reinforcement methods," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 28, pp. 338-355, 1998.

[4] E. Meng Joo and D. Chang, "Online tuning of fuzzy inference systems using dynamic fuzzy Q-learning," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, pp. 1478-1489, 2004.

[5] V. Derhami, *et al.*, "Fuzzy Sarsa Learning and the proof of existence of its stationary points," *Asian Journal of Control*, vol. 10, pp. 535-549, 2008.

[6] A. Bonarini, *et al.*, "Reinforcement distribution in fuzzy Q-learning," *Fuzzy Sets and Systems*, vol. 160, pp. 1420-1443, 2009.

[7] L. Kaelbling, *et al.*, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237-285, 1996.

[8] R. Sutton, *et al.*, "Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation," in *Proceedings of the 26th International Conference on Machine Learning*, 2009.